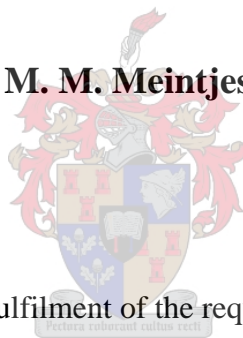


Evaluating the Properties of Sensory Tests using Computer Intensive and Biplot Methodologies

by

M. M. Meintjes



Assignment presented in partial fulfilment of the requirements for the degree of Master
of Commerce at Stellenbosch University.

Supervisors: Prof. N. J. le Roux
Dr. S. Lubbe

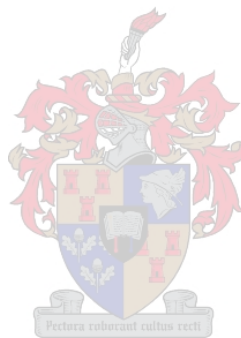
Date submitted: 6 March 2007

DECLARATION

I, the undersigned, hereby declare that the work contained in this assignment is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature:

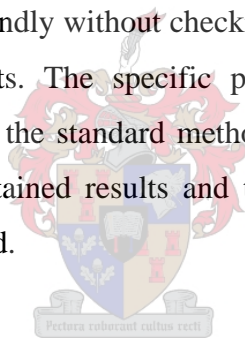
Date: 6 March 2007



SUMMARY

This study is the result of part-time work done at a product development centre. The organisation extensively makes use of trained panels in sensory trials designed to assess the quality of its product. Although standard statistical procedures are used for analysing the results arising from these trials, circumstances necessitate deviations from the prescribed protocols. Therefore the validity of conclusions drawn as a result of these testing procedures might be questionable. This assignment deals with these questions.

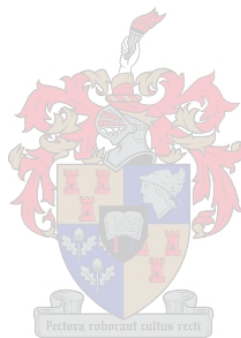
Sensory trials are vital in the development of new products, control of quality levels and the exploration of improvement in current products. Standard test procedures used to explore such questions exist but are in practice often implemented by investigators who have little or no statistical background. Thus test methods are implemented as black boxes and procedures are used blindly without checking all the appropriate assumptions and other statistical requirements. The specific product under consideration often warrants certain modifications to the standard methodology. These changes may have some unknown effect on the obtained results and therefore should be scrutinized to ensure that the results remain valid.



The aim of this study is to investigate the distribution and other characteristics of sensory data, comparing the hypothesised, observed and bootstrap distributions. Furthermore, the standard testing methods used to analyse sensory data sets will be evaluated. After comparing these methods, alternative testing methods may be introduced and then tested using newly generated data sets.

Graphical displays are also useful to get an overall impression of the data under consideration. Biplots are especially useful in the investigation of multivariate sensory data. The underlying relationships among attributes and their combined effect on the panellists' decisions can be visually investigated by constructing a biplot. Results obtained by implementing biplot methods are compared to those of sensory tests, i.e. whether a significant difference between objects will correspond to large distances between the points representing objects in the display.

In conclusion some recommendations are made as to how the organisation under consideration should implement sensory procedures in future trials. However, these proposals are preliminary and further research is necessary before final adoption. Some issues for further investigation are suggested.



OPSOMMING

Hierdie studie spruit uit deelydse werk by 'n produk-ontwikkeling-sentrum. Die organisasie maak in al hul sensoriese proewe rakende die kwaliteit van hul produkte op groot skaal gebruik van opgeleide panele. Alhoewel standaard prosedures ingespan word om die resultate te analiseer, noodsaak sekere omstandighede dat die voorgeskrewe protokol in 'n aangepaste vorm geïmplementeer word. Dié aanpassings mag meebring dat gevolgtrekkings gebaseer op resultate ongeldig is. Hierdie werkstuk ondersoek bogenoemde probleem.

Sensoriese proewe is noodsaaklik in kwaliteitbeheer, die verbetering van bestaande produkte, asook die ontwikkeling van nuwe produkte. Daar bestaan standaard toetsprosedures om vraagstukke te verken, maar dié word dikwels toegepas deur navorsers met min of geen statistiese kennis. Dit lei daartoe dat toetsprosedures blindelings geïmplementeer en resultate geïnterpreteer word sonder om die nodige aannames en ander statistiese vereistes na te gaan. Alhoewel 'n spesifieke produk die wysiging van die standaard metode kan regverdig, kan hierdie veranderinge 'n groot invloed op die resultate hê. Dus moet die geldigheid van die resultate noukeurig ondersoek word.

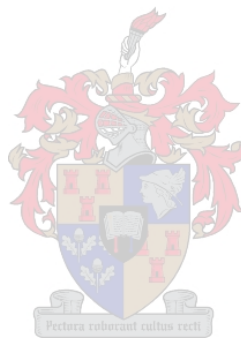


Die doel van hierdie studie is om die verdeling sowel as ander eienskappe van sensoriese data te bestudeer, deur die verdeling onder die nulhipotese sowel as die waargenome- en skoenuitverdelings te beskou. Verder geniet die standaard toetsprosedure, tans in gebruik om sensoriese data te analiseer, ook aandag. Na afloop hiervan word alternatiewe toetsprosedures voorgestel en dié geëvalueer op nuut gegenereerde datastelle.

Grafiese voorstellings is ook nuttig om 'n geheelbeeld te kry van die data onder bespreking. Bistippings is veral handig om meerdimensionele sensoriese data te bestudeer. Die onderliggende verband tussen die kenmerke van 'n produk sowel as hul gekombineerde effek op 'n paneel se besluit, kan hierdeur visueel ondersoek word. Resultate verkry in die voorstellings word vergelyk met dié van sensoriese toetsprosedures om vas te stel of statisties betekenisvolle verskille in 'n produk

korrespondeer met groot afstande tussen die relevante punte in die bistippingsvoorstelling.

Ten slotte word sekere aanbevelings rakende die implementering van sensoriese proewe in die toekoms aan die betrokke organisasie gemaak. Hierdie aanbevelings word gemaak op grond van die voorafgaande ondersoeke, maar verdere navorsing is nodig voor die finale aanvaarding daarvan. Waar moontlik, word voorstelle vir verdere ondersoeke gedoen.



ACKNOWLEDGEMENTS

I would like to thank Prof. N.J. le Roux, one of my supervisors, for all his guidance and patience throughout the completion of this assignment. His enthusiasm as well as conscientiousness is truly awe-inspiring.

Furthermore, I would like to express my appreciation to Dr. S. Lubbe, my other supervisor, who presented me with much of the necessary data and other information. Thanks for your encouragement and assistance.

I wish to thank my parents and Gerhard for their unwavering support throughout my studies; you are my pillars of strength.

Finally, a special thanks to everyone not mentioned above who played a role in my life or influenced my studies in some way during the past year.



Contents

Declaration	ii
Summary	iii
Opsomming	v
Acknowledgements	vi
1 Introduction	1
2 Brief overview of Sensory Science	5
2.1 Sensory panels and judgement criteria	5
2.2 A few considerations from sensory psychology	10
2.3 Choice of statistical test	11
2.4 Test methodologies	13
2.4.1 Background	13
2.4.2 Triangle test	13
2.4.3 Duo-Trio test	14
2.4.4 Taint test	14
2.4.5 Paired comparison testing	15
2.4.6 Binomial statistics	15
2.4.7 Chi-square tests	17
2.5 The influence of panel size	18
2.6 Replications	19
2.6.1 Influence of replicates on statistical power	22

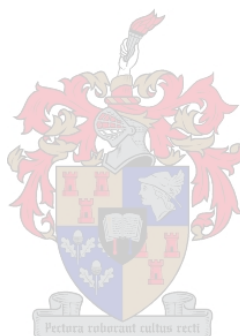


3	Implementing sensory trials in practice	25
	3.1 Application of the taint test procedure	25
	3.1.1 Description of implemented procedure	27
	3.1.2 Description of data	28
	3.2 Results obtained	30
	3.2.1 Control mean	31
	3.2.2 Treatment mean	33
	3.2.3 Test used in practice	34
	3.3 Examining the underlying distribution of allocated scores	35
	3.4 An introduction to truncated distributions	37
	3.4.1 Definition	37
	3.4.2 Estimating the parameters for the trial data	41
	3.4.3 Estimation for the coded control data	47
	3.5 Influence of panel size	52
4	Bootstrap methodology	54
	4.1 Background	54
	4.2 Application	56
	4.2.1 Statistic in question	56
	4.2.2 Bootstrap distribution of $\hat{\theta}$ for 3-day data	56
	4.2.3 Bootstrap distribution of $\hat{\theta}$ for 9-day data	57
	4.2.4 Bootstrap distribution of $\hat{\theta}$ for the combined control data	58
	4.2.5 Density estimates of Bootstrap distribution of $\hat{\theta}$	60

4.3	Bootstrap tests	63
4.4	Applying PCA biplots	70
5	Permutation tests	78
5.1	Background	78
5.2	Application	80
5.2.1	$H_{01}: \mu_{T3} = \mu_{C3}$	81
5.2.2	$H_{02}: \mu_{T9} = \mu_{C3}$	82
5.2.3	$H_{03}: \mu_{C9} = \mu_{C3}$	83
5.2.4	$H_{04}: \mu_{T9} = \mu_{C9}$	84
5.2.5	$H_{05}: \mu_{T3} = \mu_{C9}$	85
5.2.6	$H_{06}: \mu_{T9} = \mu_{T3}$	85
5.3	In conclusion	86
5.4	True permutation distribution	87
5.5	Accuracy of approximated permutation distributions	91
5.6	Association in panel scores	98
6	Comparing results from sensory tests with biplots	102
6.1	An introduction to biplots	102
6.2	Principal component analysis (PCA) biplots	103



6.3 Procedure for testing paired comparison data	104
6.3.1 Introduction to the Wilcoxon paired sample test	104
6.3.2 A normal approximation to the Wilcoxon paired sample test	105
6.4 Four standard sensory trials	105
6.4.1 Results obtained in paired discrimination tests	106
6.4.2 Results obtained in triangle tests	107
6.5 Applying biplots	108
6.6 Influence of panel size in triangle tests	111
7 Conclusion	118
Appendix	123
References	142





CHAPTER 1

INTRODUCTION

Sensory science is defined as “a scientific discipline used to evoke, measure, analyse and interpret reactions to those characteristics of foods as they are perceived by the senses of sight, smell, taste, touch and hearing” (U.S. Institute Technologists, 1975 as quoted by Stone & Sidel, 1985). The real life problems arising in consumer and sensory fields are of considerable consequence for any manufacturing, marketing and sales organisation.

For the research and development of products the recognisable attributes of a product and the influence of these on purchasing decisions of consumers are of importance. Due to the nature of these factors an interaction between food science and behavioural psychology exists, that may be studied using statistics and data analysis (Dijksterhuis, 1997). The physical composition of a product plays a decisive role in the perception of that item and therefore the chemical senses related to physiological as well as psychological aspects of a product are relevant. These aspects can be modelled mathematically and the characteristics of the resulting model evaluated.

In practice a company will have its own set of standard procedures which are often blindly implemented and analysed by their development team when modifying existing products or developing new products. There is a real danger that these procedures are implemented and analysed, without even considering whether the appropriate assumptions, on which these procedures rely, are met. Significance levels are obtained without taking graphical displays of the actual data acquired into account, and thus may be misleading. When these significance levels are used to test a given hypothesis, they may lead to faulty results. A prolonged development process due to making inadequate changes in a product will be the result of these mistakes and can be very costly. Another possibility is that a product is modified in such a way that there is a difference in taste, but this difference is not picked up in the sensory trials. If this difference is noticed by

the consumers they may switch to opposition brands which could have a serious impact on the sales of that product.

Discrimination tests are the primary concern in this study. Typically, triangle, paired comparison or duo-trio tests are implemented to address topics in this field. Due to the nature of a specific trial or preferences of an organisation, these standard procedures are often modified. The full implications of these modifications may be critical when assessing the reliability and validity of the obtained results. The data from several sensory trials, carried out at the product development centre of a specific company, were obtained. The procedures used consist of several separate discriminatory trials, carried out according to the specifications of the organisation, without a complete understanding of the underlying statistical concepts. The procedures implemented will be scrutinised, especially aspects such as the choice of test statistic and appropriate number of panellists as well as the assumed distribution with its correspondingly acquired significance levels. The data will also be analysed by implementing standard statistical tests frequently employed to test whether the means of two samples differs significantly. These results will be compared to those obtained from the organisations specified methodology typically used in its specific development process.

To further investigate the different features of the modified sensory trials, bootstrap methodology is implemented and the bootstrap distributions are employed as estimates of the underlying distributions of the test statistic being used. Kernel density estimates are also employed to obtain continuous distribution estimates. This is done separately for two similar data sets acquired from two modified, so-called taint tests, which were performed. Since the prescribed sample size seems unusually small, the effect of the sample size is also examined in some detail.

The choice of the mean as test statistic will be considered. There are other candidates for a test statistic such as the median, percentiles or even a combination of these. Looking at each of the underlying distributions separately as well as constructing a principal component analysis (PCA) biplot of all the candidate statistics in two dimensions, will cast some light on this question.

Many standard sensory tests make several distributional assumptions and thus are parametric tests. Since it is known that these assumptions are often violated when a test is carried out in practice, the parametric tests will then not be appropriate for assessing such data. Permutation tests are non-parametric and thus avoid this concern. Results from the application of permutation tests will be compared to those obtained from standard statistical, taint and bootstrap tests. Due to the small size of the data sets, the true permutation distributions are also available and are subsequently compared to the approximated permutation distribution resulting from the application of the permutation tests in practice.

To summarise, the aim of this study is to investigate the different sensory procedures implemented by a certain organisation. Several different test procedures and data will be evaluated to ascertain whether the results obtained by their prescribed procedures yield the intended results. Furthermore, multivariate techniques will also be applied to the data at hand as exploratory techniques, as well as to ascertain whether they substantiate the conclusions drawn from the univariate approach. Computer intensive methods such as the bootstrap method and permutation tests, are employed to explore the underlying distributions of the obtained data as well as the test statistics in more detail.

A brief overview of standard sensory procedures and several aspects concerned with them are discussed in Chapter 2 to gain some perspective on the topic addressed in the chapter to follow. Chapter 3 describes the standard sensory techniques as well as the modified procedure used to analyse the data under scrutiny. Here t-tests are considered to test for significant differences in the traditional way. Subsequently in Chapter 4, after a concise introduction, the bootstrap methodology will be applied to the data introduced in the previous chapter. Here the underlying distributions of several test statistics are explored. Chapter 5 contains the permutation tests which also consider the test statistics under scrutiny in the previous sections. Due to the extremely large number of permutations necessary to calculate the true distribution, approximations are usually obtained by sampling from the true distribution. In this study permutation tests are conducted in this standard way. Due to interesting features of the approximate

permutation distribution, the actual permutation distribution is calculated, which is then used to evaluate the results obtained by making use of the approximate distribution.

A different set of trials is being explored in Chapter 6. In these trials different tests are used in the development of products. The results obtained in these tests are compared to a biplot of the actual composition of the products to check whether a significant test result corresponds to a large distance between the relevant objects. The assignment is concluded in Chapter 7 where an overview of the obtained results is given together with some suggestions to the organisation where the sensory trial data were investigated.

The notation that will be used throughout is summarised in Table 1.1.

Table 1.1: Summary of necessary notation.

Symbol	Definition
$Bi(n, p)$	Binomial distribution based on n trials with the probability of success p
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\Phi_{\mu, \sigma^2}^{ABS}(y)$	Distribution function of the absolute normal distribution with mean μ and variance σ^2
$\phi_{\mu, \sigma^2}^{ABS}(x)$	Density function of the absolute normal distribution with mean μ and variance σ^2
$Z_{\alpha(2)}$	$(1 - \alpha / 2)$ 'th percentile of the standard normal distribution
H_0	Null hypothesis
\equiv	Defined as
\sim	Distributed according to
\approx	Almost equal to
\mathbf{x}	Vector of values
\bar{x}	Mean of a vector of values
$\hat{\sigma}$	An estimate of σ
PCA	Principal Component Analysis
GPA	Generalised Procrustes Analysis
ANOVA	(Univariate) Analysis of Variance

CHAPTER 2

BRIEF OVERVIEW OF SENSORY SCIENCE

Sensory science is a field of research that possesses a variety of trial procedures that may be implemented to address typical sensory questions. These methods are often implemented in a modified form, which is considered to be the standard method by the organisation that requested these trials. While the trial procedures have been modified, test procedures are frequently implemented without accounting for these changes at all. In order to judge the statistical aspects of the procedures prescribed by the organisation under consideration, basic knowledge of the sensory science is essential.

Typical sensory procedures will receive some attention in this chapter. The measurement instrument employed in sensory research consists of a panel of judges. The choice of individuals to serve on a panel, the amount of training they receive as well as the appropriate size of a panel are under consideration. The decision process underlying a conclusion drawn by an individual is of interest and some sensory psychological ideas will be discussed in order to gain insight into the topic.

How test procedures are carried out as well as the resulting statistical characteristics of the responses obtained will be discussed in some detail. The properties of the assumed distributions under the null as well as the alternative hypothesis will be discussed, since they are utilised to determine the corresponding significance level and power associated with a test. In Chapter 3 these procedures will be compared to those prescribed by the specific organisation.

2.1 SENSORY PANELS AND JUDGEMENT CRITERIA

Consumer and sensory science focuses on two aspects: the study of the product and the study of the consumer. The product may be scrutinised by considering some of its specific features or characteristics. These are the attributes which serve as the variables

and may have nominal, numerical or ordinal values. The latter two are typically obtained as the markers on a continuous range which indicate the intensity of a specific attribute that the panellist experienced. Often the number of possible markers or categories is prescribed. This raises the question whether this fixed number of categories may influence the results obtained. Optimal scaling is concerned with exactly this topic and presents methodology to obtain optimal markers used to judge a product. These markers will typically not be equally spaced as is often assumed. For example, suppose the aim is to choose six values ranging from zero to five, then the set (0.1, 0.5, 1.2, 2.5, 4, 4.8) may describe the true relationship better in terms of a statistical criterion than the obvious choice of (0, 1, 2, 3, 4, 5).

		Attributes (A_j)			
		A_1	A_2	...	A_m
Panellists ($pnls_i$)	$pnls_1$	x_{11}	x_{12}	...	x_{1m}
	$pnls_2$	x_{21}	x_{22}	...	x_{2m}
	...	x_{31}	...		
		
	$pnls_n$	x_{n1}	x_{nm}

Figure 2.1: Representation of first type of multivariate sensory data.

		Panellists ($pnls_j$)			
		$pnls_1$	$pnls_2$...	$pnls_n$
Panel(pnl_i)	pnl_1	x_{11}	x_{12}	...	x_{1m}
	pnl_2	x_{21}	x_{22}	...	x_{2m}
	pnl_3	x_{31}	...		
		
	pnl_n	x_{n1}	x_{nm}

Figure 2.2: Representation of second type of multivariate sensory data.

Data are viewed as multivariate since the following two types of data will be under scrutiny: Firstly, a single score allocated to an object by each of the judges on a panel.

The panel's scores then form one observation with each judge's score serving as a value for a specific variable, as illustrated in Figure 2.1. In the other data sets the responses will typically be in the form of scores given to certain attributes experienced by each of the judges. In Figure 2.2 (Dijksterhuis, 1997) it can be seen that the attributes serve as the variables and the judges as the samples.

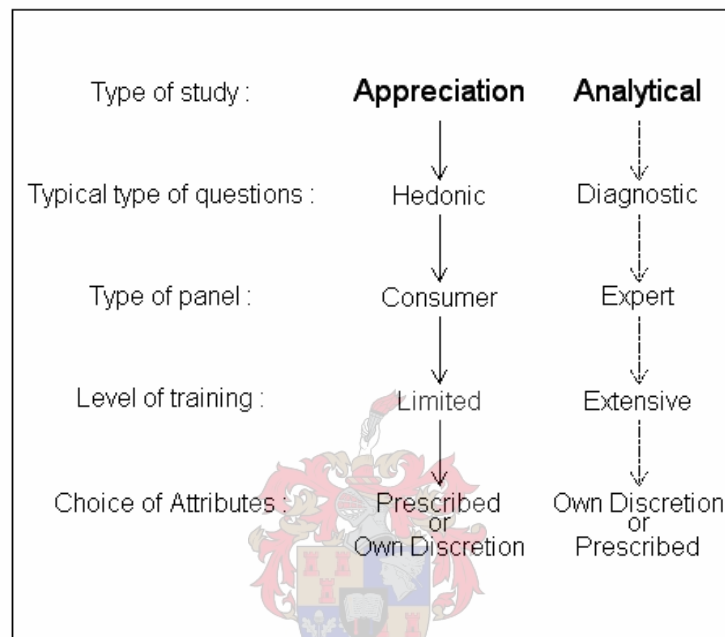


Figure 2.3: Schematic illustration of the two main types of sensory studies.

Two main types of trials exist, namely analytical and appreciation studies, which are represented in Figure 2.3. Typically, in analytical trials experts will be used to judge objects according to a specific strictly prescribed set of characteristics. Hedonic properties such as appreciation are not appropriate since the aim is not to find out whether the panel likes or dislikes the product or changes in the product, but rather to determine if changes are perceptible or what characteristics an item possesses. The panels used to answer these types of questions are extensively trained. Perception studies typically contain analytical sensory profiling which primarily employs the taste and smell senses. Here the attributes are usually not prescribed. An example of perception studies are Sensory-Instrumental studies which are chiefly analytical and examine the link between chemical (physical) and sensory aspects (Dijksterhuis, 1997).

In appreciation studies the ideal intensity of a product or preferences of consumers are under consideration thus the features under scrutiny are generally fixed and are predominantly hedonic in nature. Panels that participate in these trials consist primarily of consumers, who receive very little training. Sensory panels vary in the amount of training that they receive. Sometimes untrained individuals are selected and although investigators attempt to obtain as random a sample as possible, this is rarely completely achieved due to cost and time constraints. Thus panels usually consist of individuals selected in a quasi-random fashion that participate in studies to ascertain the reaction of the consumer. These panels are called consumer panels. Trials are performed using prescribed methodology under specified circumstances to ensure that although panellists are untrained the methodology is implemented in a consistent manner. This allows for results from separate trials to be compared to one another. When the preference testing takes place at the location where judges were recruited, the panel is referred to as a field panel.

Physiological differences between panellists will cause the intensity experienced by them to vary, while a lack of vocabulary may lead to a different classification of the same sensation. The latter of these may be resolved by training the panel when analytical problems are under consideration. If this training is effective the differences in responses may be assumed to be due to individual differences and should be removed effectively by the correct standardisation and may consequently be analysed using PCA methodology. While PCA methodology will be discussed in Chapter 6 and it will also be used in a specialised application in Chapters 5 and 6, the above-mentioned special application of PCA falls beyond the scope of this study.

If however, psychological differences have an effect as well, more complicated methods are necessary to correct for this interpretation-effect. Since the panel serves as the measuring instrument in sensory research, these differences correspond to different calibrations of the measuring instruments. Consequently, an individual's perception of intensity is comparable between objects but not between respective panellists. For this reason the values of a judge should be viewed as a set and using generalised Procrustes

analysis (GPA) these sets may be transformed to agree optimally, even if the number or choice of attributes differs.

Another way of discriminating between the various kinds of panels is by looking at the kind of questions that they are asked to resolve. Intensively trained panels will typically consider analytical topics while untrained panels are generally used to determine the appreciation of an object and thus deal with hedonic questions such as preferences and enjoyment. Therefore, the questions faced by a panel will become more analytical the more training they received prior to the trial. Figure 2.4 gives a schematic representation of the continuous range of training received by panellists (Dijksterhuis, 1997, Figure 3).

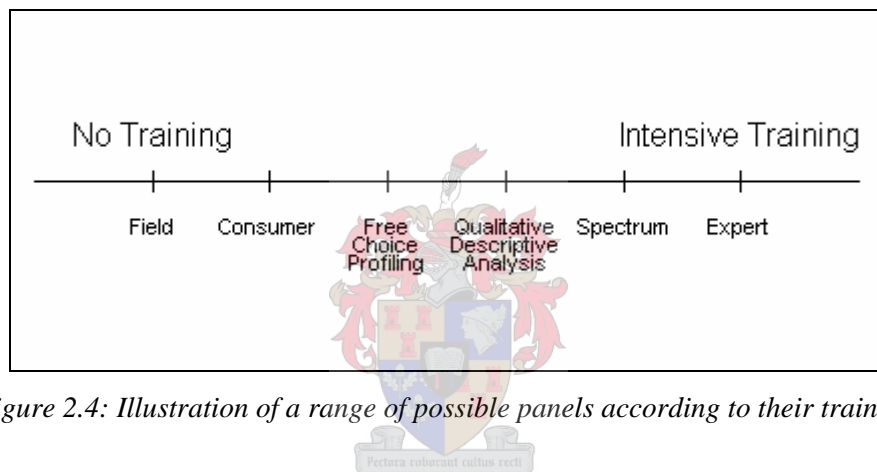


Figure 2.4: Illustration of a range of possible panels according to their training.

Profiling can be described as the process of describing certain features of an object. Conventionally, the vocabulary employed to do this is prescribed to the panellists prior to the trial, which can be considered as the process of calibrating the panel. Examples of these are Spectrum and Qualitative Descriptive Analysis (QDA) panels, as displayed in Figure 2.4. According to Dijksterhuis (1997) vocabulary of descriptive attributes are formed for both these two types of panels. Spectrum and QDA panels vary, however, in the amount of as well as the methods of training. Calibrating is performed in order to minimise the variation in responses obtained from the panel for identical objects; thus minimising the error.

Free choice profiling (FCP) panels differ according to the criteria used to judge an object. These panels do not examine an object with regard to a fixed set of attributes, since each panellist is allowed to develop her/his own list of characteristics with which

to judge the object. Panels with this quality are typically not extensively trained. The fact that each panellist may compile her/his own list of attributes complicates the analysis of the data since averages for each individual cannot be calculated due to the fact that the summing of different attributes does not make sense. Individual difference models or GPA have to be used to analyse such data.

2.2 A FEW CONSIDERATIONS FROM SENSORY PSYCHOLOGY

Since the panellists serve as the measuring instruments, understanding the mental processes leading to a conclusion could be of great help when trying to attain some insight into the sensory procedure being analysed. To gain a perspective on these thought processes, some basic considerations from the field of experimental psychology will now be introduced.

According to Osgood (1953), for a stimulus to be detectable by the nervous system an absolute minimum threshold must be exceeded, thus it has to be of at least some minimal magnitude. The receptors only register a change in the environment when the energy applied to them surpasses some lower threshold. This implies that although two objects may be physically or chemically different, the mind will only register this difference if it is adequately large. It gets more complicated, however, since the minimal threshold that needs to be exceeded in order for the mind to recognise the difference, may differ between individuals and is also not even constant within an individual.

A panellist may perceive the same stimulus at separate points in time differently due to fatigue or changes in her/his physiological composition or state of mind. In taste for instance, the receptors quickly adapt to a certain level of a stimulus and one would therefore require increasing levels of a stimulus in order for it to be still detected, which Osgood (1953) refers to as adaptation. Physiological differences refer to the physiological aspects of an individual and how these aspects affect the way they experience a situation. Alpern, Lawrence & Wolsk (1967) state that the diversity in receptor cells situated on the tongue, for example, will influence an individual's sensitivity to taste. Several sources of variation in the perception of stimuli exist. One of these is differences in the sensitivity of the senses of individuals. Another factor that

according to Woodworth & Schlosberg (1960) may have an effect on taste is how large an area was subjected to the stimulus.

When considering all of these influences that affect results, the validity of sensory tests are brought into question. An important question to consider is whether the aim of these procedures is to measure true differences between objects or whether to measure the *perceived* difference. In most sensory trials the latter of these are at issue and adjustment methods are typically implemented to learn exactly how large a stimulus has to be for an individual to perceive the difference.

2.3 CHOICE OF STATISTICAL TEST

There are several factors that influence the choice of statistical test for different sensory evaluation methods. One of these is the assumptions regarding the family of distributions from which the data originate. If these assumptions are not met the results obtained by application of the test methodology may be rendered meaningless. Parametric or non-parametric methods may be employed based on the measurement level and the nature of the population distribution. Whether the observations are dependent or independent is also an important issue to take into account. This is crucial since there may be several sources of dependency that the investigator is unaware of such as the correlation between results obtained by using the same panellists.

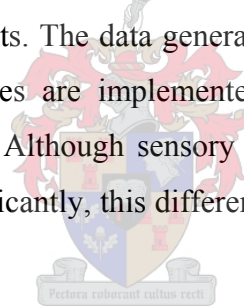
Often the choice of test is limited due to the nature of the data generated by a given sensory technique such as discrimination testing or ranking. Bower (1996) states that frequently used test methodologies include:

1. One sample test: where the sample results are compared to a known population parameter.
2. Two sample test: if the number of treatments compared is limited to two, typically the following would be implemented:
 - two sample test (independent samples)
 - paired test (related samples)
3. Analysis of variance (ANOVA) is used when three or more treatments are considered.

An important issue to keep in mind when implementing single factor ANOVA, is that it fails to recognise related samples. However this can be addressed by using a block design and thus including a block effect for panellists in the model.

Another factor to consider is whether the statistical procedures should be used to analyse the objects or the panellists. When the sensory panel is used as a measuring instrument to judge some characteristic of a product, the results will be generalised to the larger population of products. In other studies the reaction/preference of the consumer is under consideration and the panel is a random sample of consumers that generates information about the population of consumers. The specific aim of each study should therefore be kept in mind since it could cause confusion in inexperienced users.

Discrimination trials aim to determine whether a significant difference exists between objects or characteristics of objects. The data generated are usually nominal or ordinal. Non-parametric testing procedures are implemented since no assumption as to the distribution of the data is made. Although sensory discrimination procedures indicate whether two samples differ significantly, this difference is not quantified which restricts the usefulness of the results.



To avoid confusion regarding the purpose and conclusion of significance tests, it is recommended that the objective, the obtained result as well as the null hypothesis and the alternative hypothesis should be stated clearly before the sensory procedure is executed (Bower, 1996). The significance level as well as the scope of the test should also be stated. The significance level is the probability of obtaining at least such an extreme result if the null hypothesis is true. This is also the type I error (α), i.e. the probability of rejecting the null hypothesis when in fact it is true. As this error is to be minimized, significance levels smaller than α are required. It is however not that simple, since the smaller the value of α the more difficult it is to reject the null hypothesis when the alternative hypothesis is true, which leaves one with a test with very little power. The issue of the power associated with a specific test procedure will be discussed in Section 2.6. When the hypotheses are specified, it is important to

consider whether a one-sided or two-sided alternative hypothesis is relevant. To ensure that the statistical assumptions are met, the conditions under which the trial is performed are also of great consequence.

2.4 TEST METHODOLOGIES

2.4.1 Background

Several standard sensory tests exist, for which the complete procedure and distributional assumptions are available. These tests are implemented in sensory discrimination trials carried out by a sensory analyst, who typically has a very restricted knowledge of the application of statistical principles in practice. Since these tests are often modified to address some practical issue, the distributional assumptions and requirements of the original procedure such as independence are often not met. To evaluate these changes and the way they may impact obtained results, the standard procedures and their properties will first be discussed.

2.4.2 Triangle test

The triangle test is implemented in sensory discrimination trials to check whether there is a discernable difference between two objects. A judge is presented with three coded samples (say X123, Y456 and Z789). The panel is told that two of these are the same and that one is different. They are asked to identify the one which differs from the others. Either two of the samples will be from the control type and one (Z789, say) a test sample; or two samples will be from the test products and the other one from the control. This is known to the investigator but not to the panel. The hypothesis of interest thus is:

$$H_0: \text{Z789 does not differ from the others} \quad \text{vs} \quad H_a: \text{Z789 differs from the rest}$$

Now the variable of interest is defined as $X \equiv$ number of judges identifying Z789 correctly (i.e. the number of successes). In Section 2.4.6 it is explained why this procedure implies that $X \sim Bi(n, 1/3)$, under the null hypothesis, if there were n judges participating in the trial.

2.4.3 Duo-Trio test

Duo-trio tests deal also with sensory discrimination. The duo-trio test is very similar to the triangle test in that 3 objects are considered. Each judge is presented with a control item that serves as a reference (say X123) and is presented with two coded samples (Y456 and Z789, say). Panellists are instructed to identify the coded sample that is the same as the reference. It is known by the trial analyst which of these items are control items and which are not. The hypothesis is:

$$H_0: Y456 = Z789 \quad \text{vs} \quad H_a: Y456 \neq Z789$$

The number of judges who are able to identify the treatment objects correctly, is known as the number of successes and as before the binomial distribution applies under the null hypothesis. However, if there is no difference between the control and the treatment, the judge will have to guess which one of the two items is the test item. Therefore the probability of a fixed number of successes amongst n judges is distributed according to $Bi(n, 1/2)$ distribution.

It is important to note that the decision processes underlying assessors' judgements in the triangle and duo-trio tests differ. When asked which object differs from the rest in a triangle test, a judge is seeking to find a difference and this may cause bias in their judgement because they can become oversensitive to perceiving differences which are imaginary or insignificant. In duo-trio tests, seeking to determine which objects are similar, a judge aims to perceive similarities which is not the same as trying to recognise differences.

2.4.4 Taint test

Taint tests are discrimination tests applied to determine whether a specific treatment may cause a significant taint or change in the taste of an item. This procedure is a combination of a duo-trio and a paired comparison test. Each panellist is presented with a reference sample as well as two coded samples. One of these coded samples is a treated sample and the other a control sample, but the panellist is unaware of this. The panellist is instructed to award a score to each of the coded samples quantifying how

much they differ from the reference. These scores have a fixed range, for example 1–9, where a low score corresponds with a small difference and a high score with a large difference.

2.4.5 Paired comparison testing

These tests are implemented in profiling as well as in discrimination studies. A judge is presented with two items. Typically one of these would be an original and the other a modified version of a specific product. The judge is either supplied with a set of attributes or requested to form her/his own. They are requested to evaluate each object on all of these attributes and award a value to each attribute. These values will usually be scores on a fixed scale from 0 to 5. When there is no evidence to support a normality assumption, Wilcoxon paired sample tests are applied to these scores to determine whether a significant difference exists between the two objects for each attribute. If the normality assumption holds, the paired t-test can be implemented.

2.4.6 Binomial statistics

Binomial statistics are typically implemented in discrimination tests. The binomial distribution is discrete and deals with the probability of a given number of successes in a fixed number of independent trials. Thus this distribution may be applied to sensory discrimination, where the number of panellists serve as the fixed number of trials. In discrimination tests (except in the duo-trio tests), individuals are typically required to identify the object that differs from the rest, thus the perceived rather than the physical difference is of importance. If a triangle test is performed, for instance, the panellist identifying the correct object is seen as a success. The null hypothesis is that there is no difference amongst the three items under consideration. If this is true, in order for the trial to be successful, the panellist must guess the correct item which has a $1/3$ chance of happening. Therefore under the null hypothesis the number of positive identifications of the objects is distributed according to $Bi(n, 1/3)$, where n is the number of panellists involved in the study. It is obvious that giving the panel an option of saying that none of the objects differ, will modify the null hypothesis which illustrates the influence of variations in the standard procedures.

The observed significance level is the probability that at least the observed number of successes occurred under the null distribution, i.e. that the panellist could not detect any difference but, being forced to make a choice, chooses an object at random. The lower the probability, the more unlikely it is that such an event occurred by chance and therefore the conclusion is made (if the observed significance level is sufficiently small) that there is a significant difference between the objects. The observed significance level is compared to the already specified α -level and if the observed significance level is equal to or smaller, the null hypothesis is rejected in favour of the alternative hypothesis. Thus the test statistic of interest here is the observed significance level.

From this it is apparent that although the test does indicate whether there is a significant difference between objects, it does not provide any information as to in what manner objects differ or quantify the size of this perceived difference.

When comparing different types of sensory discrimination procedures, such as the triangle and duo-trio tests, the null distributions differ and thus fewer successes may be needed in the triangle tests to obtain a significant result than in the duo-trio tests. Bower (1996) states that a trade-off exists between the apparent advantage of the triangle test and the confusion that may be experienced by the panellists. In triangle tests the panellists are required to identify the sample that differs from the rest, thus they have to compare all three with each other. A duo-trio test only requires the panellists to compare two samples to a reference, thus they are not required to compare them with each other, and therefore fewer comparisons are required than in the triangle test. Therefore the probabilistic advantage of the triangle test should not be confused with the relative sensitivity of the test since the decision processes differ considerably.

Another aspect to consider is whether an insignificant test result implies that there is no difference between a test and control object. Similarity testing is implemented to determine whether two objects are the same. Since the decision process underlying a similarity judgement differs from that underlying the judgement of whether there is a difference between objects, results obtained in similarity testing will not necessarily correspond with those obtained in difference testing. The probability of a type II error

(i.e. the chance that two objects are declared identical if they actually differ) is of special interest in similarity testing.

2.4.7 Chi-square tests

Discrimination data lend themselves to the implementation of chi-square tests where the observed frequency is compared to the expected frequency. Once again this procedure is applied under the assumption that there is no difference between samples. Now the number of possible categories allowed as response, may be two or more and a continuity correction needs to be made if there are only two categories (such as success or failure). It is also recommended that there should preferably be 40 or more panellists with no less than five incidents per category (Bower, 1996).

The χ^2 -statistic is evaluated under the assumption that the null hypothesis is valid by implementing the following formula (if there are more than two categories), (Bower, 1996):

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad \text{where } O \text{ is the observed frequency per category,}$$

E is the expected frequency per category,
and the summation is taken over all categories.

In the case where there are only two categories (for example success or failure), application of the continuity correction leads to the following result (Bower, 1996):

$$\chi^2 = \frac{(\text{Absolute difference between sample selections} - 1)^2}{(\text{total number of selections})} .$$

The test statistic of interest here is the evaluated χ^2 -statistic which is compared to the tabulated χ^2 -percentile with $(k - 1)$ degrees of freedom, if there are k categories. For example in a duo-trio test, the expected split frequency under the null hypothesis is 50:50. Therefore, say there is a panel of size $n = 20$, then the expected frequency is 10 for each category.

Table 2.1: Possible contingency table for duo-trio example.

	Actual number of successes	Actual number of failures	Row Total
Expected number of successes	XXXXXXX	XXXXXXX	10
Expected number of failures	XXXXXXX	XXXXXXX	10
Column Total	8	12	20

Table 2.1 gives some indication as to how the observed and actual observations may be represented, where the specific cell frequencies are replaced by XXXXXXX since they do not play a role in computing the test statistic. If the observed frequencies are 8 and 12, respectively, this leads to a χ^2 -statistic of:

$$\chi^2 = \frac{(|8 - 12| - 1)^2}{20} = 0.45.$$

In the one-tailed test, this is compared to the $\chi^2_{1,0.05} = 2.71$. Thus the null hypothesis that the two categories are the same cannot be rejected at a 5% significance level.

When comparing the binomial test with the uncorrected chi-square test, it can be shown that the uncorrected chi-square test is more powerful, and therefore it has a higher probability of rejecting the null hypothesis. This is however only true in the uncorrected case and does not apply in the adjusted case. Bower (1996) states that although there is only a small difference between these two tests, the risk (β) of accepting a false null hypothesis is less for the uncorrected chi-square test. If the results obtained by implementing both these tests differ, further testing is necessary to verify which result is correct. The statistical power of tests, specifically when replications are included in difference tests, will be referred to in Section 2.6.

2.5 THE INFLUENCE OF PANEL SIZE

An important aspect to consider is that of the appropriate panel size and how this may affect results. The larger the panel, the smaller the frequency of agreement required to obtain a significant result. For paired comparison testing it is recommended by the British Standards Institution (1982) that a minimum of 7 expert panellists or 20

consumer panellists are needed to attain reliable results. These tests compare a test item and a control item to decide whether a detectable difference exists.

Complications may arise when the panel is large, as that may inflate the sources of error, for example trial preparation and data transcribing errors. A further side effect is that a frequency that is only slightly larger than expected from chance, may now be deemed significant and thus marginal differences are detected. This fact may undermine the practical value of the result. This shortcoming may be addressed by decreasing the α -level. In similarity testing a larger panel is required to minimise the β risk. Since the required panel size may become unrealistically large when both the type I and type II errors are being minimised, the investigator has to trade-off the statistical issues, the practicality of requiring a large panel and the costs involved.

2.6 REPLICATIONS

Replications in sensory tests can complicate the analyses of the responses. Typically, replications consist of performing several separate trials using the same panel. In order to combine data from similar trials, independence assumptions are necessary which often do not hold. If different assessors are used in the two separate trials, it is expected that the same results would be obtained if the experiment is carried out under the same circumstances. Therefore, the panellist effect can be considered to be random, similar to the approach followed in the ANOVA approach to sensory profiling where the block effect for the specific panel is included in the model and this panel effect is considered random.

Brockhoff (1995) suggested the above-mentioned random effect approach to sensory replication and evaluated several methods that may be utilised, including the overdispersion method. The overdispersion method simply amounts to correcting the appropriate totals and then using these corrected totals in standard testing procedures. This method was further researched by Brockhoff & Schlich (1998) who proposed the following:

If n = number of assessors
 k = number of replicates per assessor
 x = total number of successes;

then the ‘overdispersion’ parameter $\hat{\sigma}^2$ is essentially given by the variance of the individual frequencies of successes for the n assessors corrected for non-zero chance probabilities. A success may be whatever is considered to be the correct response. The ‘corrected number of observations’ is calculated as $nk / \hat{\sigma}^2$ and the ‘corrected number of successes’ is given by $x / \hat{\sigma}^2$. Then the testing and power calculations are performed as usual using the corrected statistics above.

Brockhoff & Schlich (1998, equations 1–5) provide closed form expressions used to implement the method described above to estimate the overdispersion. There are basically only four possible outcomes when considering the overdispersion: the panels differ and a) there is a difference between the products, or b) there is no difference between the products. On the other hand the panels may be similar and c) the products differ, or d) the products do not differ. It is important to note that if there is no difference in the product, the judges have to give homogeneous responses.



The most common reason for making use of replicates is that not enough panellists are available, while larger panels are required to obtain a more powerful test. Compensating for this shortage of panellists by making use of replicates is a simple solution according to Brockhoff & Schlich (1998) who explains the procedure as follows:

1. The objective of a test should be defined clearly, which includes choosing the appropriate levels of α and β as well as the percentage of distinguishers (above chance) which should be detected with probability $1 - \beta$. Distinguishers are defined as panellists that are able to discriminate between samples, if a difference exists, with a probability larger than that expected from chance.
2. Consult Schlich’s table (Schlich, 1993) to determine the minimum total number of respondents required (denoted N).

3. An assumption on the level of overdispersion to be expected in the population of respondents being sampled has to be made. This $\hat{\sigma}^2$ must be between 1 (homogeneity) and k (maximal heterogeneity). Since k is not known in advance, the expected overdispersion can be expressed as a proportion q between 1 and k : $\hat{\sigma}^2 = (1-q) + qk$. So $q = 0$ corresponds to homogeneity ($\hat{\sigma}^2 = 1$), $q = 1$ corresponds to maximal heterogeneity ($\hat{\sigma}^2 = k$). Here q is attained from past experience. If no prior knowledge of the level of heterogeneity is available, a value of 1/3 is suggested, which empirical studies according to Brockhoff & Schlich (1998) have shown to be the most realistic.
4. Now n and k are chosen to ensure that $nk/(1-q+qk) \geq N$.
5. If the number of panellists n is given, step 4 implies that k should be chosen as:

$$k = \frac{N(1-q)}{n - qN}.$$

6. Therefore, when $q = 1/3$ it follows that:

$$k = \frac{2N}{3n - N}.$$

or the smallest integer larger than the computed value. A negative value for k , that is if $n < qN$, corresponds to the situation where the required risks set earlier cannot be met. The degree of heterogeneity in such data, sets a limit to the possible levels of risks obtainable that even an infinite number of replications would not remedy.

7. When the number of replications k is known in advance, step 4 implies that n should be chosen as:

$$n = N(q + (1-q)/k)$$

Now for $q = 1/3$ it follows that $n = \frac{N(k+2)}{3k}$

or the smallest integer larger than the computed value. This always gives a possible value for n .

The methods discussed above are straightforward to implement but are hampered by the assumption that the decision process of each assessor functions independently within each replication. Fatigue or educated learning is ignored which may be ill-advised.

Another effect that is not taken into account is the effect of a specific session, when the number of replications required necessitates more than one trial session. Psychometric considerations should be investigated in these cases and their effects on the validity of the replication approach evaluated.

2.6.1 Influence of replicates on statistical power

Replicated binomial tests are often found where n panellists perform k replicates of the difference test. The statistical power of these tests is now under consideration. Overdispersion (Brockhoff & Schlich, 1998) and several models such as the beta-binomial distribution (Ennis & Bi, 1998) and generalised linear models (Hunter, Piggot & Lee, 2000) have been suggested to deal with this issue. Brockhoff (2003) notes that the problem with implementing overdispersion is that there is no formal model defined, and argues that straightforward application of the two models mentioned above is not viable since the individual probabilities have a fixed lower limit. He further suggests that corrected versions of the beta-binomial and generalised linear model will be more appropriate to account for this inadequacy.

The probability that a discrimination test will detect a difference under the alternative hypothesis, i.e. when there is a significant difference between objects, is called the power of the procedure. An nk -binomial test is often implemented and consists of carrying out k replicate trials with a panel of size n . The main reason for applying the nk -binomial distribution is that, despite the replications, the binomial distribution still holds under the null hypothesis. The power of the α -level, nk -binomial test may be defined as:

$$\text{Power} \equiv P_{H_a} (X \geq x_{critical})$$

where $x_{critical}$ is the $(1-\alpha)$ th percentile of the binomial distribution. This holds independently of assumptions concerning the replications. For power calculations the distribution under the alternative hypothesis is of interest, however, and a model for this distribution is required. The probability distribution for the total number of correct answers X when there is in fact a sensory difference (i.e. under the alternative hypothesis) needs to be specified. Thus the distribution which generates the probability

that x successes are measured, given that there is a difference between subjects needs to be modelled.

As mentioned above, Brockhoff (2003) suggests the following possible models: beta-binomial, generalised linear mixed model (GLMM), mixture of two binomials, corrected beta-binomial and the corrected generalised linear mixed model. The differences between these models lie in the different distributions used to model the random panellist effect. Each of these models assumes the binomial model for each panellist and models the random panellist effect by some distribution. Although all of these models are dependent on the same number of parameters (i.e. two), these parameters cannot be compared directly and thus are unable to serve as a simple way of choosing the most suitable model. These models can only be fitted to data containing some heterogeneity which restricts the number of situations in which they can be applied. Thus when implementing this, the overdispersion should first be calculated and if this is more than one, the above-mentioned models can be fitted. If this constraint is not met Brockhoff (2003) states that the most suitable alternative model is the nk -binomial and there is no (nontrivial) distribution to estimate. In the examples explored by Brockhoff it is apparent that the distributions for the probability of a success by an individual fitted by the beta-binomial and the GLMM-models are very similar.

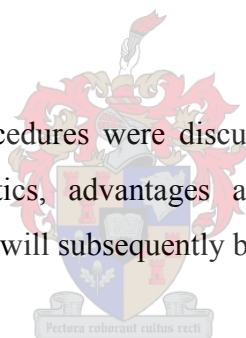
Power estimates are calculated by making use of Monte Carlo methods and a complete algorithm is contained in Brockhoff (2003). The relationship between the power of the test and the number of replications is of interest to the sensory investigator. Is the power of a test with a few panellists carrying out a large number of replications the same as that of a large panel carrying out only a few replications? Although there is a connection between the heterogeneity of the panel and the achieved power, it has been shown that the loss in power due to the lack of homogeneity in the panel, has very little effect on the power. Tables for the power of duo-trio and triangle tests are available (Brockhoff (2003): Table 3 and 4) for various levels of n and k . It has been shown that if there is a medium effect, i.e. a medium difference exists between the objects under consideration, and a fixed panel, a few replications will vastly improve the power of a test. An interesting fact mentioned in Brockhoff (2003) is that in some cases a small panel

completing five replications of a test will have greater power than that of a fixed large panel carrying out a single trial, even if the total number of observations is the same. This may be attributed to the fact that substituting assessors by replicates, although the power is unchanged, has costs in terms of precision and an increase in variability.

These results however do not imply that a few panellists completing a large number of replications should be used rather than a larger panel, with each of these judges only performing a few replications. One disadvantage of the former is that training is not taken into account, i.e. if a panellist is confronted with the same taste repeatedly, it may become easier for her/him to discern. Another issue is fatigue, which may cause the judgement of the panel to deteriorate. An additional drawback is that, despite the fact that a test may possess more power, it does not mean that it will necessarily be able to quantify differences better as the result of an increase in the variability of the data.

In conclusion

In this chapter standard test procedures were discussed to gain some perspective on sensory trials, their characteristics, advantages and disadvantages. A specialised application of sensory procedures will subsequently be discussed.



CHAPTER 3

IMPLEMENTING SENSORY TRIALS IN PRACTICE

Sensory trials are carried out in the development of most consumable products. Although standard sensory methods exist, companies often have their own prescribed method for implementing these procedures. The data that are considered in this assignment comes from an organisation that implements a specific variation of taint tests. Due to confidentiality the name of this organisation will not be disclosed and will be referred to as “the organisation” in this assignment. The taint test procedure is also a standard procedure and is implemented as specified in the MQM protocol, contained in Figure 3.1. This procedure is performed at one of the product development centres by sensory analysts with limited statistical knowledge. Questions arose, concerning the accuracy and obtained significance levels of results, which led to this study. In this chapter the results obtained by the organisation’s test procedures will be compared to those typically performed to test for significance differences in samples.

3.1 Application of the taint test procedure

The taint test procedure described in the MQM protocol was not implemented by the organisation exactly as prescribed in Figure 3.1, although the organisation regards the MQM protocol as a standard procedure for examining whether there is a perceivable taint in a product and will be using it in all their taint tests. The sensory procedure that was performed in the two trials under consideration, will now be described and will be referred to as the implemented MQM (IMQM) protocol. Following this description, two data sets resulting from a variation of the previously mentioned protocol are investigated in detail.

Taint Test

The test protocol below may be followed, or alternatively that detailed within EN 1230-2:2001. When reporting results, reference MUST be made to the test protocol followed.

MQM 1.2 (Sept 02)

Equipment

Kilner jar

Milk chocolate

Aluminium foil

Frequency

As required

Method

1. Rinse out jar with distilled water and leave to air-dry.
2. Cut A4 sample of material under test into 10mm strips
3. Place 10mm strips into the jar with a foil cup containing 5 pieces of chocolate and replace the lid.
4. Wrap another 5 or 6 pieces in foil and place in an empty, taint free jar. This is the control sample.
5. Leave the test and control jar at room temperature for a minimum of 24 hours.
6. When the 24 hours has elapsed, eat a piece of chocolate from the control sample. Rinse our mouth with water.
7. Eat a piece of chocolate from the test jar and compare its taint with the one from the control sample.
8. Results are assessed as follows:

0	No taint
1	Barely perceptible taint
2	Perceptible taint
3	Moderate taint
4	Strong taint
5	Very strong taint
9. The test must be carried out by a panel of preferably 5 people. However, a minimum of 3 people is permissible.
10. An average of the panels' results is taken. The figure is recorded, along with any comments, against the production batch reference.

Figure 3.1: The document describing the MQM protocol for implementing taint tests in detail.

The following terminology will be used to describe the taint tests and the obtained data: Control items are standard objects that were not subjected to any treatment and thus serve as a reference to determine whether the treatment had any perceivable effect on the product. Test items, also referred to as treatment items, are those objects that did receive the treatment in question. In the case to be considered the control objects are pieces of chocolate and the test objects are pieces of the same chocolate that were placed with product packaging. The aim of this trial is to test whether product packaging may cause a taint in the taste of the product.

3.1.1 Description of implemented procedure

The aim of this trial is to determine whether printed product packaging material would cause a noticeable taint in the taste of chocolates. The expectation of the sensory analyst is that relatively freshly printed packaging material (three days after printing) might cause some taint while “airing” the packaging material for a longer period (i.e. nine days) might allow enough time for the evaporation of printing solvent residues so that no taint can be perceived.

Chocolates were placed with product packaging as described in Figure 3.1. Each judge was then presented with a control piece of chocolate as well as two labelled pieces, one of which came from the control block and the other from the tainted chocolate. Note that typically when implementing the MQM protocol, there would only be one test and one control piece, but in this specific IMQM procedure the labelled control piece was added to get some idea as to how accurate the panel is. The panel is not aware that one of the labelled pieces to be judged is a control piece.

The judges were asked to first taste the control, rinse their mouths and then taste a labelled piece. The labelled piece is to be compared to the control piece and a value between zero and five assigned to the labelled piece reflecting how much this piece differs from the control (zero indicating that there is no difference and five that there is an unmistakable difference). The judges are allowed to give integer as well as half values and thus allowing scores to range from zero to five, give results on a nine point scale as is typically used in paired comparison testing. After assigning a value, this

process is repeated, tasting the control piece again, rinsing the mouth and then tasting the second labelled piece, assigning a value between zero and five.

The mean of the allocated scores for the coded test sample is computed and compared to a critical value of one as specified by the MQM protocol. If the mean exceeds this critical value, the null hypothesis that there is no difference between the reference and the coded test sample is rejected and the conclusion is made that the treatment did cause a significant taint in the chocolates.

3.1.2 Description of data

The IMQM protocol described in Section 3.1.1 was performed on two separate occasions using product packaging that was printed three and nine days earlier. Expert panels consisting of nine judges participated in both these trials. The data values measured for all nine judges are contained in Table 3.1 and will be referred to as the 3-day data in what follows.

The second data set was obtained in exactly the same way as described above, but now the chocolates were placed with packaging nine days after the printing process and once again only nine judges participated in the trial. The 9-day data, contained in Table 3.2, consists of the recorded values for the test and control chocolate pieces. The judges received specific identity numbers, and as shown in the tables displayed below, some of the judges participated in both trials.

Table 3.1.: 3-day data: Data obtained in taint test.

Panellist	Score: Test Block	Score: Control Block	$y_i \equiv$ Test score – Control score
i	2	0	2
ii	3	1	2
iii	1	0	1
iv	4	0	4
v	2	1	1
vi	0	0	-0.5
vii	0	0	0
viii	2	0	2
ix	3	0	3
\bar{x}	1.8889	0.2778	1.6111
$s\hat{e}_{\bar{x}} = s / \sqrt{n}$	0.4547	0.1470	0.4698

Table 3.2.: 9-day data: Data obtained in taint test.

Panellist	Score: Test Block	Score: Control Block	$y_i \equiv$ Test score – Control score
i	2	1	1
iii	2	1	1
iv	0	0	0
vii	0	0	0
viii	1	0	1
ix	2	0	2
x	1	0	1
xi	1	2	-1
xii	2	0	2
\bar{x}	1.2222	0.4444	0.7778
$s\hat{e}_{\bar{x}} = s / \sqrt{n}$	0.2778	0.2422	0.3239

Allocating scores to quantify the perceived difference between objects is not the same than simply having to discriminate whether there is a difference. Moreover allocating scores also influences the decision process underlying the panellists' judgements. Thus it can be assumed that the decision process differs from that of the standard duo-trio procedure since a score has to be allocated. It cannot be assumed that every score larger than zero would have been noted as a difference if the question had been whether there was a difference between objects or not. Due to this fact, the binomial distribution is not an appropriate distribution from which to obtain the test statistic. The results obtained with the 3-day data set and the 9-day data set may also not be independent from each other since some of the judges were involved in both trials. These aspects will be addressed in Section 5.6.

Standard statistical tests for deciding whether two objects differ statistically will subsequently be implemented. The assumptions underlying these tests will also be considered. Following this, the IMQM test procedure for this trial will be examined. The obtained significance levels as well as the power of the test procedure will be evaluated and results compared to those obtained by a newly modified testing method.

These statistical properties of the tests can be investigated parametrically by making some distributional assumptions. Alternatively, a non-parametric computer intensive approach can be followed by considering bootstrap as well as permutation testing procedures. The former of these procedures will also be used to examine in more detail the distribution of the test statistic, used in earlier procedures.

3.2 RESULTS OBTAINED

The Student's t-test is typically implemented by statisticians to determine whether two means differ significantly. In order to implement this procedure it is assumed that the difference between the coded test sample and the reference comes from a $N(\mu_T, \sigma_T^2)$ distribution. Similarly, the distribution of the difference between the reference and the coded control sample is assumed to be a $N(\mu_C, \sigma_C^2)$ distribution. Since the coded control samples are compared to a control sample as reference, it is expected that the distribution of the difference between the coded piece tested (which actually comes

from the control sample) and that of the reference piece should yield $\mu_C = 0$. Further μ_T represents the mean difference between the test and reference, experienced by the panellist.

3.2.1 Control mean

The assumption of normality is made to allow the implementation of Student's t-test. The IMQM test procedure used in practice was developed for $\hat{\theta} = \bar{x}$, the mean of the observed data points. When testing whether the mean of the control pieces differs significantly from θ_0 , the appropriate test statistic is:

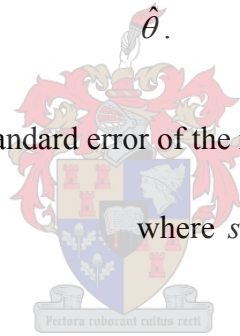
$$T_0 = \frac{\hat{\theta} - \theta_0}{s\hat{e}_\theta}, \quad \text{where } \theta_0 = 0 \text{ under the null distribution}$$

and $s\hat{e}_\theta$ is the estimated standard error for

$\hat{\theta}$.

The equation for estimating the standard error of the mean is

$$s\hat{e}_{\bar{x}} = \frac{s}{\sqrt{n}}, \quad \text{where } s = \left\{ \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \right\}^{1/2}. \quad (3.1)$$



Thus for the data obtained for the 3-day trial the hypotheses under consideration are,

$$H_0 : \mu_{C3} = 0 \quad \text{vs} \quad H_a : \mu_{C3} > 0$$

and a test value of:

$$T_0 = \frac{0.278 - 0}{0.147} \approx 1.8898$$

is obtained. This corresponds to a calculated significance level of 0.0477 when compared to the Student's t-distribution with eight degrees of freedom. Therefore the null hypothesis that there is no difference between the control pieces can be rejected at a prescribed 5% significance level. This result is somewhat alarming since if a panel is so hypersensitive that they perceive a difference between two control pieces, they should always perceive a significant difference between test and control items, rendering the result to be meaningless. Keep in mind that the panellists are not aware that one of the

coded samples is drawn from the control samples. The hypersensitivity can be attributed to the fact that if an individual is told to quantify *the difference* between two objects, she/he looks to find any difference and thus will notice any trivial difference.

Similarly, when the test statistic for the control data of the 9-day data is analysed to test the following hypothesis:

$$H_0 : \mu_{C9} = 0 \quad \text{vs} \quad H_a : \mu_{C9} > 0 ,$$

the subsequent result is obtained:

$$T_0 = \frac{0.444 - 0}{0.2422} \approx 1.8353 .$$

This is associated with a p-value of 0.0519 with eight degrees of freedom. Here the null hypothesis that the control pieces do not differ significantly from zero cannot be rejected at a prescribed 5% significance level. Although the null hypothesis cannot be rejected, the obtained significance level is still very small.

Since both the control samples originate from the same underlying distribution, the scores can be pooled to test whether the mean of the underlying distribution differs significantly from zero. Thus the hypotheses considered are:

$$H_0 : \mu_C = 0 \quad \text{vs} \quad H_a : \mu_C > 0 .$$

The appropriate value of the test statistic is:

$$T_0 = \frac{0.3611 - 0}{0.1389} \approx 2.6000 .$$

An associated p-value of 0.0093 is obtained by using a t-distribution with 17 degrees of freedom. Therefore the null hypothesis that the mean of the distribution is equal to zero can be rejected at a 5% significance level.

These results bring accuracy of the panel into doubt, since differences were perceived between two samples which both came from the control chocolates. On the other hand, results obtained by implementing the t-test should be interpreted carefully since the normality assumption might be questioned.

3.2.2 Treatment mean

Typically the investigator will perform a paired t-test to ascertain whether there is a significant difference between the control and treated items. This procedure entails defining a new variable which is the difference between the score for the treated item and that of the control item. These variables now become the observations under scrutiny, and the aim is to test whether the mean of these new observations differs significantly from zero.

Define: $x_{ci} \equiv$ score allocated to the control piece by the i-th judge
 $x_{ti} \equiv$ score allocated to the treated piece by the i-th judge.

Then: $y_i = x_{ti} - x_{ci}$, the difference between the two scores allocated by the i-th judge. Now the appropriate one sample test may be performed, as above, with the differences y_1, \dots, y_n replacing the scores previously used.

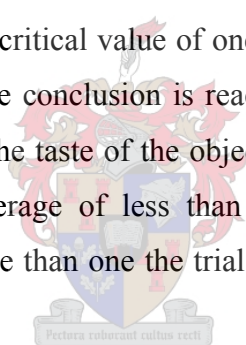
In the case of the 3-day data, the mean difference is 1.611 with a standard error of 0.4698. These values lead to $T_0 \approx 3.4296$, corresponding to a p-value of 0.0045 with eight degrees of freedom, thus the hypothesis that the difference between the scores for the test and the control items is zero is rejected at a 1% significance level. Due to lack of compliance to the previously mentioned assumptions associated with this test methodology, this result may not be deemed trustworthy. This result will also have to be compared to that found by the permutation testing procedure.

Similarly, the mean difference for the 9-day data is 0.7778 with a standard error of 0.3239, which corresponds to $T_0 \approx 2.4010$ associated with a p-value of 0.0216 with eight degrees of freedom. Once again the null hypothesis is rejected at a 5% significance level. These results were to be expected since the mean of the control data differed significantly from zero. This brings into question whether the procedure of measuring the difference between the treatment and control is effective.

When ignoring the control data and the fact that the data do not comply with the normality assumptions, a t-test can be performed to test whether the mean of the test data differs significantly from zero. This is done simply for exploratory purposes. The test statistic for the 3-day test scores is 4.1538, which renders a significance level of 0.0016 and thus the mean of the 3-day test data differs significantly from zero, as was expected. Similarly, a test statistic of 4.4 is obtained for the 9-day test scores. This corresponds with a significance level of 0.0011 and consequently it is concluded that the mean of the 9-day data also differs significantly from zero.

3.2.3 Test used in practice

In the previous sections it is shown how to analyse data obtained in taint trials utilising statistical principles. In practice, however, the MQM procedure prescribed by the organisation under consideration amounts to calculating the mean of the treatment scores and comparing this with a critical value of one. When this value is exceeded the null hypothesis is rejected and the conclusion is reached that the treatment did have a statistically significant effect on the taste of the object. This standard procedure further specifies that even when an average of less than one is obtained, but any of the panellists awarded a score of more than one the trial is to be repeated. This is however not implemented.



Furthermore the scores awarded to the control block are not taken into account at all, since these scores would typically not be available. This implies that the sensitivity or possible oversensitivity of the panel is ignored, which could have a substantial influence on the final decision.

When implementing the above prescribed procedure a mean of 1.8889 for the 3-day and 1.2222 for the 9-day data were obtained. These values correspond to a significant difference between the test and control data in the 3-day as well as the 9-day trial since the value of one is exceeded. However, this procedure does not take the variation of the data into account which might be expected to have some effect on the obtained significance level.

3.3 EXAMINING THE UNDERLYING DISTRIBUTION OF ALLOCATED SCORES

The data under consideration consist of scores allocated to two coded samples. It is known by the sensory analyst that one of these samples comes from the test chocolates and the other from the control chocolates. When performing the trial, the panellists are instructed to award a score to each sample quantifying the perceived difference between the coded sample and the reference. This value however only indicates the size of the difference and not the direction. If for instance it was the sweetness levels that differed, the test items could either be sweeter or less sweet. If the test item is less sweet a negative value for the difference would naturally be awarded, but now only the magnitude of this difference is quantified. Therefore, the scores awarded are actually the absolute values of the perceived differences.

Figures 3.2 and 3.3 contain the histograms of the allocated test scores for the 3-day and 9-day trials. The scores awarded are allowed to range from zero to five, although no score larger than four was awarded. It is assumed that the underlying distribution of the difference between the coded test sample and reference sample is distributed according to a $N(\mu_T, \sigma_T^2)$ distribution. Similarly, the underlying distribution of the difference between the coded control sample and the reference sample is assumed to be the $N(\mu_C, \sigma_C^2)$ distribution. Note that these parametric assumptions are made with regards to the distribution of the perceived differences and not the allocated scores, thus negative values can be observed.

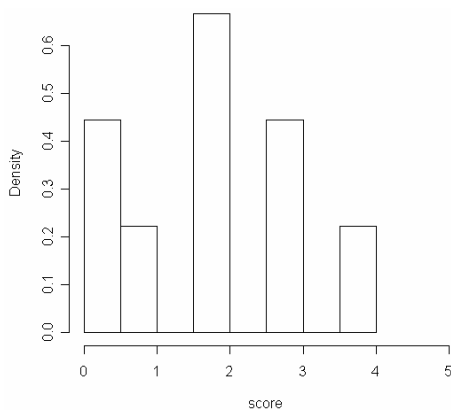


Figure 3.2: Histogram of test scores for the 3-day trial.

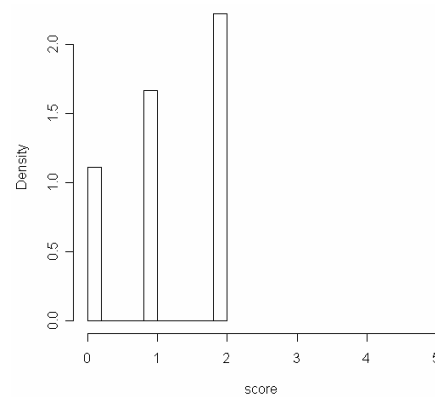


Figure 3.3: Histogram of test scores for the 9-day trial.

The perceived differences are not measured, however. Only the size of this difference is quantified by the awarded scores. Thus a score is the absolute value of the difference experienced by the panellist. The relationship between the observed difference and the allocated scores can be described as follows:

Let : $D \equiv$ the perceived difference between the coded sample and the reference,

$X \equiv$ the score awarded by the panellist .

Therefore: $X = |D|$ with $D \sim N(\mu, \sigma^2)$

and

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(|D| \leq x) = P(-x \leq D \leq x) = P(D \leq x) - P(D \leq -x) \\ &= \Phi_{\mu, \sigma^2}(x) - \Phi_{\mu, \sigma^2}(-x) \\ &= \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy - \int_{-\infty}^{-x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy. \end{aligned}$$

Further:

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) \\ &= \frac{d}{dx} \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy - \frac{d}{dx} \int_{-\infty}^{-x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\ &= \frac{d}{dx} \int_{-\infty}^x \phi_{\mu, \sigma^2}(y) dy - \frac{d}{dx} \int_{-\infty}^{-x} \phi_{\mu, \sigma^2}(y) dy \\ &= \phi_{\mu, \sigma^2}(x) \frac{d}{dx}(x) - \phi_{\mu, \sigma^2}(-x) \frac{d}{dx}(-x) \\ &= \phi_{\mu, \sigma^2}(x) + \phi_{\mu, \sigma^2}(-x) \end{aligned} \tag{3.2}$$

where Φ_{μ, σ^2} and ϕ_{μ, σ^2} is the distribution and density function of the $N(\mu, \sigma^2)$ distribution. Now the absolute-normal distribution is defined as follows:

Let, $Y \sim N(\mu, \sigma^2)$ and $X = |Y|$

$$\text{then: } \Phi_{\mu, \sigma^2}^{ABS}(y) = P(Y \leq y) = \begin{cases} \Phi_{\mu, \sigma^2}(y) - \Phi_{\mu, \sigma^2}(-y), & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore the underlying distribution of the scores allocated to a sample is an absolute normal distribution. The underlying distribution for the *difference* between the coded test sample and the reference is $N(\mu_T, \sigma_T^2)$ and the corresponding distribution for the *difference* between the coded control sample and the reference is $N(\mu_C, \sigma_C^2)$. Although scores of larger than five were not permitted in the trial procedure, the absolute normal distribution does not take this into account. This issue will be addressed by considering truncated distributions

3.4 AN INTRODUCTION TO TRUNCATED DISTRIBUTIONS

In the previous sections it was assumed that the scores were distributed according to an absolute normal distribution. This does not take into account the effect that the scores are restricted to the interval $[0; 5]$, i.e. according to the previous absolute normality assumptions values of higher than five can be obtained but such values are not permitted in the trial procedure implemented. Another shortcoming is that the absolute normal distribution is continuous. This implies that a panellist has the freedom to give any numerical value between zero and five. For simplicity the IMQM procedure does not allow panellists to assign values such as 1.5225. The underlying continuous distribution is thus measured in a discrete fashion. Therefore discretisation of the obtained distribution may be necessary after fitting the appropriate distribution that complies with the above-mentioned assumptions.

3.4.1 Definition: Truncated Distributions

Truncated distributions may be used to address the fact that the absolute normal distribution does not account for the fact that scores of higher than five were not allowed. In these distributions, the random variable has a known distribution which is only valid within a restricted range of values. Often a well-known distribution is restricted to a fixed interval within its usual range. In this case the distribution that needs to be restricted to the interval $[0; 5]$ is the absolute normal distribution. A formal definition of the density function of a truncated continuous random variable, as defined by Mood, Graybill & Boes (1963), is as follows:

Let X be a random variable with density function $f_X(x)$ and cumulative distribution function $F_X(x)$. Further, let $I_{[a,b]}(x)$ be the indicator function such that:

$$I_{[a,b]}(x) = \begin{cases} 1, & \text{if } x \in [a, b] \\ 0, & \text{otherwise.} \end{cases}$$

Now the density of X truncated at a and b may be written as:

$$f_{X,a,b}(x) = \frac{f_X(x)I_{[a,b]}(x)}{F_X(b) - F_X(a)}.$$

For the problem under consideration, the absolute normal distribution is truncated at zero and five. Note that the absolute normal distribution already implies that all the probability is associated with non-negative values. Therefore truncating the absolute normal distribution *at* five is equivalent to truncating the distribution *between* zero and five. Therefore the appropriate distribution for which parameter estimates are required has the form:

$$f_{X,0,5}(x) = \frac{\phi_{\mu,\sigma^2}^{ABS}(x)I_{[0,5]}(x)}{\Phi_{\mu,\sigma^2}^{ABS}(5) - \Phi_{\mu,\sigma^2}^{ABS}(0)} \quad (3.3)$$

where $\phi_{\mu,\sigma^2}^{ABS}(x)$ is the density function defined in equation (3.2) corresponding to the $\Phi_{\mu,\sigma^2}^{ABS}(y)$ distribution function. To estimate the parameters μ and σ^2 , the control and test data sets will be used. Using the parameter estimates significance levels are obtained and these can be compared to those based on the critical value of one used in IMQM.

Maximum likelihood estimation is typically implemented to obtain parameter estimates for a specific distribution. This consists of maximizing the probability of obtaining the observed sample from that specific family of distributions. The maximum likelihood

function consists of the product of the density functions, evaluated for each observation. By taking the partial derivatives of the logarithm of the likelihood functions with respect to the parameters and equating these to zero, the estimating equations are obtained. Since closed form equations for the maximum likelihood estimates (MLE) are not available, in this case numerical maximisation is used to obtain appropriate parameter estimates. The derivations of the appropriate equations for the truncated absolute normal equations are as displayed below.

$$L = \prod_{i=1}^n f_{X,0,5}(x_i) = \prod_{i=1}^n \frac{\phi_{\mu,\sigma^2}^{ABS}(x_i) I_{[0,5]}(x_i)}{\Phi_{\mu,\sigma^2}^{ABS}(5) - \Phi_{\mu,\sigma^2}^{ABS}(0)}$$

$$= \prod_{i=1}^n \frac{\phi_{\mu,\sigma^2}^{ABS}(x_i)}{\Phi_{\mu,\sigma^2}^{ABS}(5)}, \quad \text{since } x_i \in [0,5], \forall i$$

and $\Phi_{\mu,\sigma^2}^{ABS}(0) = 0 \forall \mu, \sigma^2$

but $\Phi_{\mu,\sigma^2}^{ABS}(x) = \Phi_{\mu,\sigma^2}(x) - \Phi_{\mu,\sigma^2}(-x),$ since $x_i \geq 0, \forall i$

and $\phi_{\mu,\sigma^2}^{ABS}(x) = \phi_{\mu,\sigma^2}(x) + \phi_{\mu,\sigma^2}(-x)$

thus $L = \prod_{i=1}^n \frac{\{\phi_{\mu,\sigma^2}(x_i) + \phi_{\mu,\sigma^2}(-x_i)\}}{\{\Phi_{\mu,\sigma^2}(5) - \Phi_{\mu,\sigma^2}(-5)\}}.$

Therefore the expressions that need to be maximised are:

$$\ln(L) = \sum_{i=1}^n [\ln\{\phi_{\mu,\sigma^2}(x_i) + \phi_{\mu,\sigma^2}(-x_i)\} - \ln\{\Phi_{\mu,\sigma^2}(5) - \Phi_{\mu,\sigma^2}(-5)\}] \quad (3.4)$$

Taking the derivatives with respect to μ and σ^2 , and equating these to 0 give:

$$\sum_{i=1}^n \frac{d}{d\mu} [\ln\{\phi_{\mu,\sigma^2}(x_i) + \phi_{\mu,\sigma^2}(-x_i)\} - \ln\{\Phi_{\mu,\sigma^2}(5) - \Phi_{\mu,\sigma^2}(-5)\}] = 0 \quad (3.5)$$

$$\sum_{i=1}^n \frac{d}{d\sigma^2} [\ln\{\phi_{\mu,\sigma^2}(x_i) + \phi_{\mu,\sigma^2}(-x_i)\} - \ln\{\Phi_{\mu,\sigma^2}(5) - \Phi_{\mu,\sigma^2}(-5)\}] = 0. \quad (3.6)$$

Since expressions (3.5) and (3.6) do not readily lead to simple closed form expressions for μ and σ^2 , numerical maximisation of equation (3.4) was carried out by writing the following R-code using the R-function **optim()** to minimise the negative of equation (3.4) as follows:

```

dabsnorm<-function(y,mu,sigma2)
{ifelse(y<0,0,dnorm(y,mu,sqrt(sigma2))+dnorm(-y,mu,sqrt(sigma2)))}

pabsnorm<-function(y,mu,sigma2)
{ifelse(y<0,0,pnorm(y,mu,sqrt(sigma2))-pnorm(-y,mu,sqrt(sigma2)))}

optim(par=c(mu,sigma2),fn=function(x){a<-x[1];b<-x[2];-sum (log(
dabsnorm(datvec,a,b) / (pabsnorm(5,a,b)))}),
lower=0.1,upper=25,method="L-BFGS-B")
or
method= "Nelder-Mead"
with the appropriate changes in the lower and upper arguments.

```

When both of these parameters were estimated¹, values of (1.8684, 2.1471) and (1.2039, 0.6617) for the test data of the 3-day and 9-day data sets, respectively, were obtained. Figures 3.4 and 3.5 display the histogram of these two data sets with the truncated normal distribution based on these parameter estimates superimposed. These fitted distributions will subsequently be used to estimate the power of the sensory test implemented. Although these estimates seem to fit the histogram reasonably well for a continuous distribution, the inadequacy of fitting a continuous distribution to discrete scores is clearly shown.

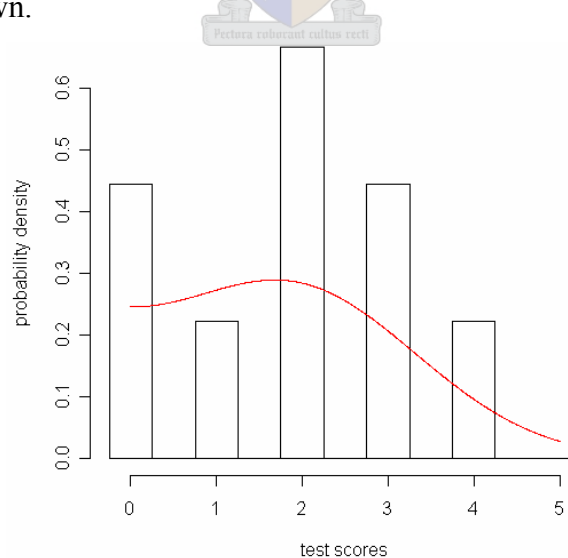


Figure 3.4: Histogram of 3-day test data with truncated density function (red line) superimposed.

¹ These estimates were obtained by implementing the R-function **MLE.estimate()** on the appropriate data set with default methods. This function is contained in the Appendix.

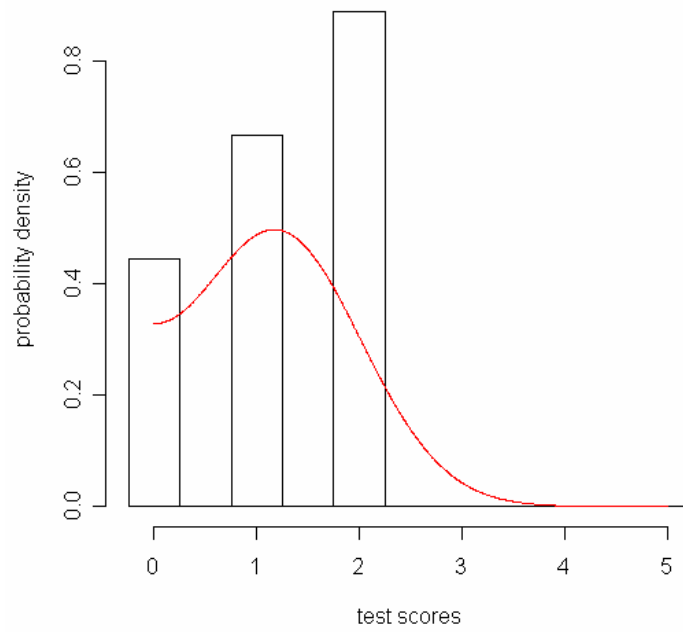


Figure 3.5: Histogram of 9-day test data with truncated density function (red line) superimposed.

3.4.2 Estimating the parameters for the trial data

Under the null hypothesis for the 3-day and 9-day trials it is assumed that the truncated absolute normal distribution with parameters $(0, \sigma_0^2)$ is valid for the test data, since there is assumed to be no difference between the coded test sample and the reference sample. Under the assumption that the data are distributed according to a truncated normal distribution as explained above, it is necessary to estimate σ_0^2 for the 3-day and 9-day test data sets with μ held constant at 0 to be able to calculate the observed significance level. The maximum likelihood estimates¹ for σ_0^2 for the test data of the 3-day and 9-day data are 7.3022 and 2.1275, respectively. Therefore choosing a fixed critical value such as one does not make much sense since this value will correspond to different significance levels depending on the estimate $\hat{\sigma}_0^2$.

When not assuming that $\mu = 0$, thus not reducing the model, maximum likelihood estimates are calculated for both μ and σ . The estimated density functions for this model (red) are superimposed on that of the reduced models (blue) and the histograms. These are shown in Figures 3.6 and 3.7 for the two data sets.

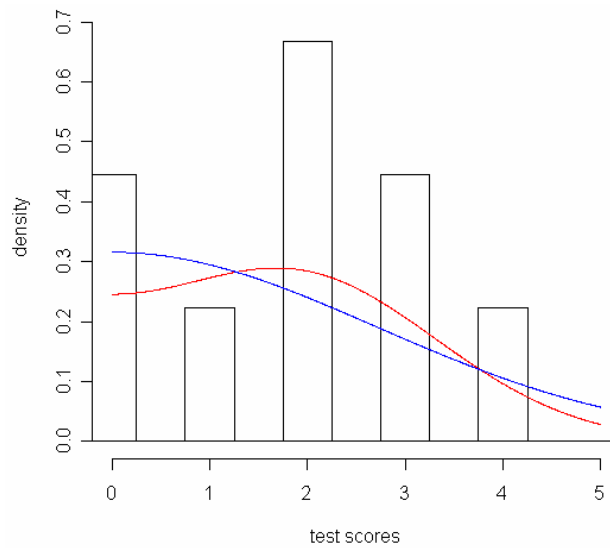


Figure 3.6: Histogram of 3-day test data with the truncated density function for the distribution under the null hypothesis (red line) and the complete model (blue line).

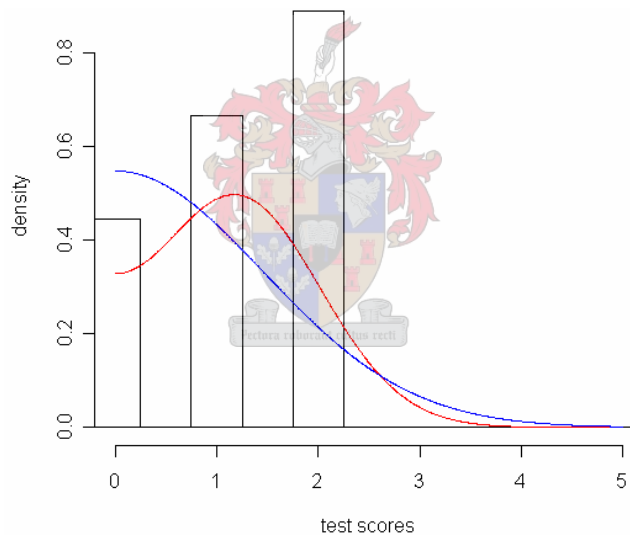


Figure 3.7: Histogram of 9-day test data with the truncated density function for the distribution under the null hypothesis (red line) and the complete model (blue line).

There is only a slight difference between the two models in Figure 3.6 but there are substantial differences in the shape of the distributions in Figure 3.7 as well as between the distributions shown in Figure 3.7 and those in Figure 3.6. This illustrates the dangers of using a fixed critical value.

To calculate the significance level of the prescribed procedure, the cumulative distribution of a truncated normal distribution, with lower bound 0 and upper bound 5, will have to be evaluated for $x = 1$ and parameters $(0, \hat{\sigma}_0)$. This is now done.

$$\begin{aligned}
 \alpha_{observed} &= 1 - \int_0^1 f_{X,0,5}(x) dx \\
 &= 1 - \frac{\int_0^1 \phi_{\mu, \sigma^2}^{ABS}(x) dx}{\Phi_{\mu, \sigma^2}^{ABS}(5) - \Phi_{\mu, \sigma^2}^{ABS}(0)}, \quad \text{from (3.3)} \\
 &= 1 - \frac{\Phi_{\mu, \sigma^2}^{ABS}(1) - \Phi_{\mu, \sigma^2}^{ABS}(0)}{\Phi_{\mu, \sigma^2}^{ABS}(5) - \Phi_{\mu, \sigma^2}^{ABS}(0)} \quad (3.6)
 \end{aligned}$$

$$= 1 - \frac{\Phi_{\mu, \sigma^2}^{ABS}(1)}{\Phi_{\mu, \sigma^2}^{ABS}(5)}, \quad \text{since } \Phi_{\mu, \sigma^2}^{ABS}(0) = 0 \forall \mu, \sigma^2. \quad (3.7)$$

The significance levels² attained in the two testing procedures considered are therefore 0.6915 and 0.4927, respectively. These values are extremely large and it shows that the probability of rejecting the null hypothesis, when it is actually true, is so large that the null hypothesis cannot be rejected based on this critical value of one.

The corresponding significance levels² for the test statistics (the mean of the test scores) of 1.8889 and 1.2222 are 0.4491 and 0.4017, thus the null hypothesis cannot be rejected. Figures 3.8 and 3.9 display the extremely large probability of a type I error when a fixed value of one is used in the case of the 3-day test data and the 9-day test data, respectively. These results are not deemed to be accurate due to the fact that the truncated distribution does not fit the data adequately and the test statistics used do not take the variation contained in the test scores into account.

² These significance levels were obtained by implementing the R-function `shaded.area.graph()` for $\mu = 0$ and appropriate estimates of $\hat{\sigma}_0^2$ and critical value.

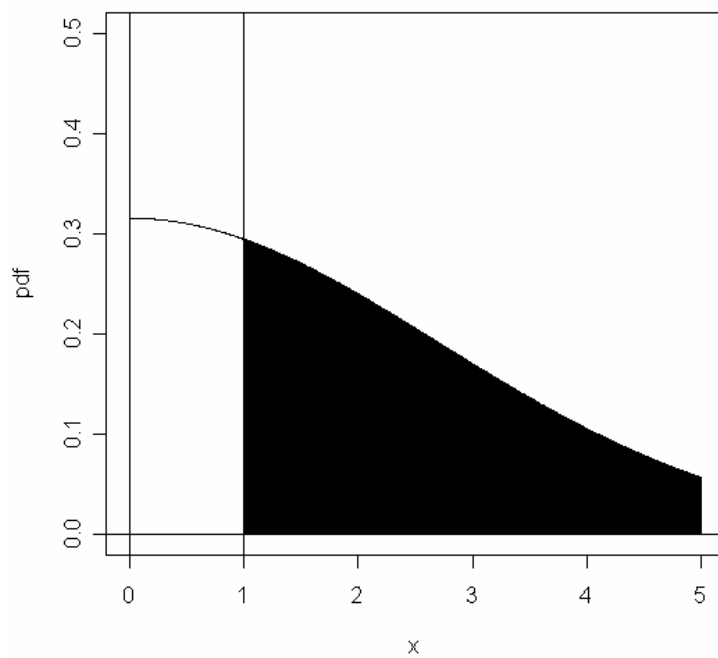


Figure 3.8: The truncated density function for the 3-day test data under the null hypothesis to illustrate the inadequate significance level for a critical value of one.

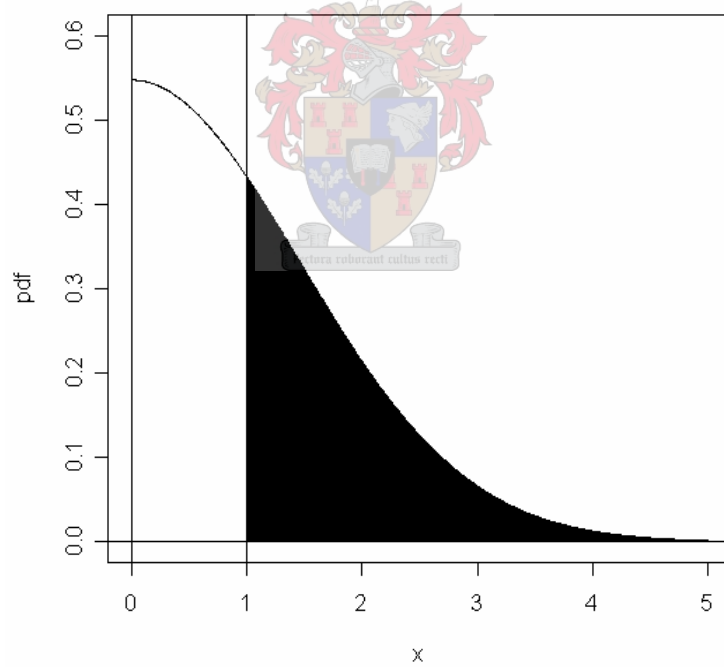


Figure 3.9: The truncated density function for the 9-day test data under the null hypothesis to illustrate the inadequate significance level for a critical value of one.

In order to calculate the obtained probability of a type II error for either trial the following needs to be done: Since the probability of a type II error is the probability of not rejecting the null hypothesis when the alternative hypothesis is true, it will be

necessary to obtain the ML-estimates for the parameters of the alternative model and use them to calculate the probability that the mean of the observed values would exceed the fixed value of one which is prescribed in the MQM protocol. The ML-estimates were obtained in Section 3.4.2. The cumulative distribution function is evaluated for the parameter estimates of (1.8684, 2.1471) and (1.2039, 0.6617), respectively. The probability of a type II error² for the two trials is 0.2557 for the 3-day data and 0.3977 for the 9-day data, respectively.

Figures 3.10 and 3.11 illustrate the shape of the distributions under the alternative hypothesis for the two testing procedures considered. The shaded area in each graph denotes the power of the test. These figures illustrate the substantial variation in the shape of the alternative hypothesis and therefore the probability of a type II error will also vary substantially.

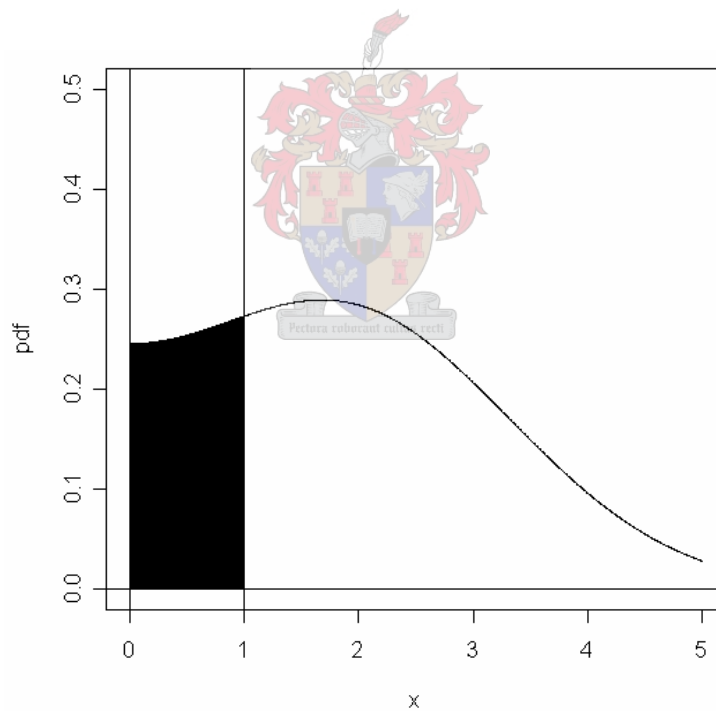


Figure 3.10: The truncated density function for the 3-day test data for the distribution under the alternative hypothesis. The shaded area denotes the probability of a type II error when using a critical value of one.

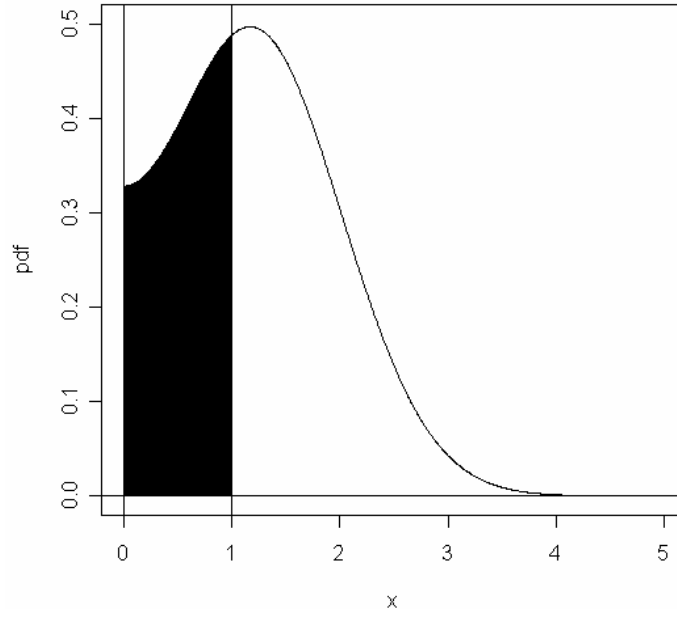


Figure 3.11: The truncated density function for the 9-day test data for the distribution under the alternative hypothesis. The shaded area denotes the probability of a type II error when using a critical value of one.

When using the estimated truncated distribution functions for the 3-day and 9-day data, respectively, the appropriate critical values³ to obtain a 5% significance level under the null hypothesis are 4.3060 for the 3-day data and 2.8516 for the 9-day data, respectively. These values are estimated by considering equation (3.6):

$$\alpha = 1 - \int_0^{x_{CRIT}} f_{X,0,5}(x) dx = 1 - \frac{\Phi_{\mu,\sigma^2}^{ABS}(x_{CRIT})}{\Phi_{\mu,\sigma^2}^{ABS}(5)}$$

$$1 - \alpha = \frac{\Phi_{\mu,\sigma^2}^{ABS}(x_{CRIT})}{\Phi_{\mu,\sigma^2}^{ABS}(5)}.$$

Therefore:

$$\Phi_{\hat{\mu},\hat{\sigma}^2}^{ABS}(x_{CRIT}) = (1 - \alpha) \Phi_{\hat{\mu},\hat{\sigma}^2}^{ABS}(5)$$

$$x_{CRIT} = \Phi_{\hat{\mu},\hat{\sigma}^2}^{ABS^{-1}}\{(1 - \alpha) \Phi_{\hat{\mu},\hat{\sigma}^2}^{ABS}(5)\}$$

³ These estimates for the critical value were obtained by implementing the R-function `crit.val.trunc.abs.norm()` for $\mu = 0$ and the appropriate $\hat{\sigma}_0^2$ estimate, contained in the appendix.

$$\begin{aligned}\Phi_{\hat{\mu}, \hat{\sigma}^2}^{ABS}(x) &= \Phi_{\hat{\mu}, \hat{\sigma}^2}(x) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(-x) = p_x \\ \Phi_{\hat{\mu}, \hat{\sigma}^2}^{ABS}(x_{CRIT}) &= (1 - \alpha)p_5.\end{aligned}\tag{3.7}$$

When combining the test data from both these trials, a critical value of 3.7456 is needed to obtain a 5% significance level. It thus follows that the variation in the data and the corresponding changes in the shape of the estimated density function will have a large influence on the obtained significance level. Furthermore, it can be seen that the critical values as well as the power of the testing procedure can vary substantially although all of them show that a critical value of one is too small. In order to determine an appropriate critical value a more extensive study is recommended. It may be necessary to develop a discrete form of the truncated distribution to obtain more accurate results.

3.4.3 Estimation for the coded control data

The problem when considering the test data of the two trials, is that the correct result, i.e. whether a significant difference between the test and reference control sample exists or not, is unknown. It is known however that the coded control samples and the reference are samples from the same underlying distribution, and thus no significant difference exist between them. Therefore the truncated absolute normal distribution is now used to estimate the distribution under the null and alternative hypotheses for the 3-day and 9-day trials as well as for the combined data set. The control data can be combined in order to get a larger sample for estimating the distribution, since both control samples come from the same underlying distribution.

Under the null hypothesis $\mu = 0$ and therefore the underlying distribution only depends on σ_0^2 . The maximisation function used to estimate the parameters for the test data makes use of the method of Byrd et. al. (1995) which allows the user to prescribe upper and lower bounds for the parameters. This is done to ensure that permissible estimators are obtained. For the control data, another method is used, due to the fact that the mean of the data is so small that the usual estimating procedure just returns the lower bound as an estimate. This unconstrained method rendered very similar estimates for the parameters for the test data and thus seems reliable.

When implementing this procedure for the control data, the following estimates¹ for σ_0^2 were obtained: 0.2500 for the 3-day, 0.6666 for the 9-day and 0.4583 for the combined control scores. Figures 3.11 a) – c) contain the histograms and corresponding fitted truncated absolute normal distributions under the null hypothesis. Although the estimates of σ_0^2 as well as the histograms vary slightly, the shape of the fitted truncated absolute normal distribution remains approximately the same.

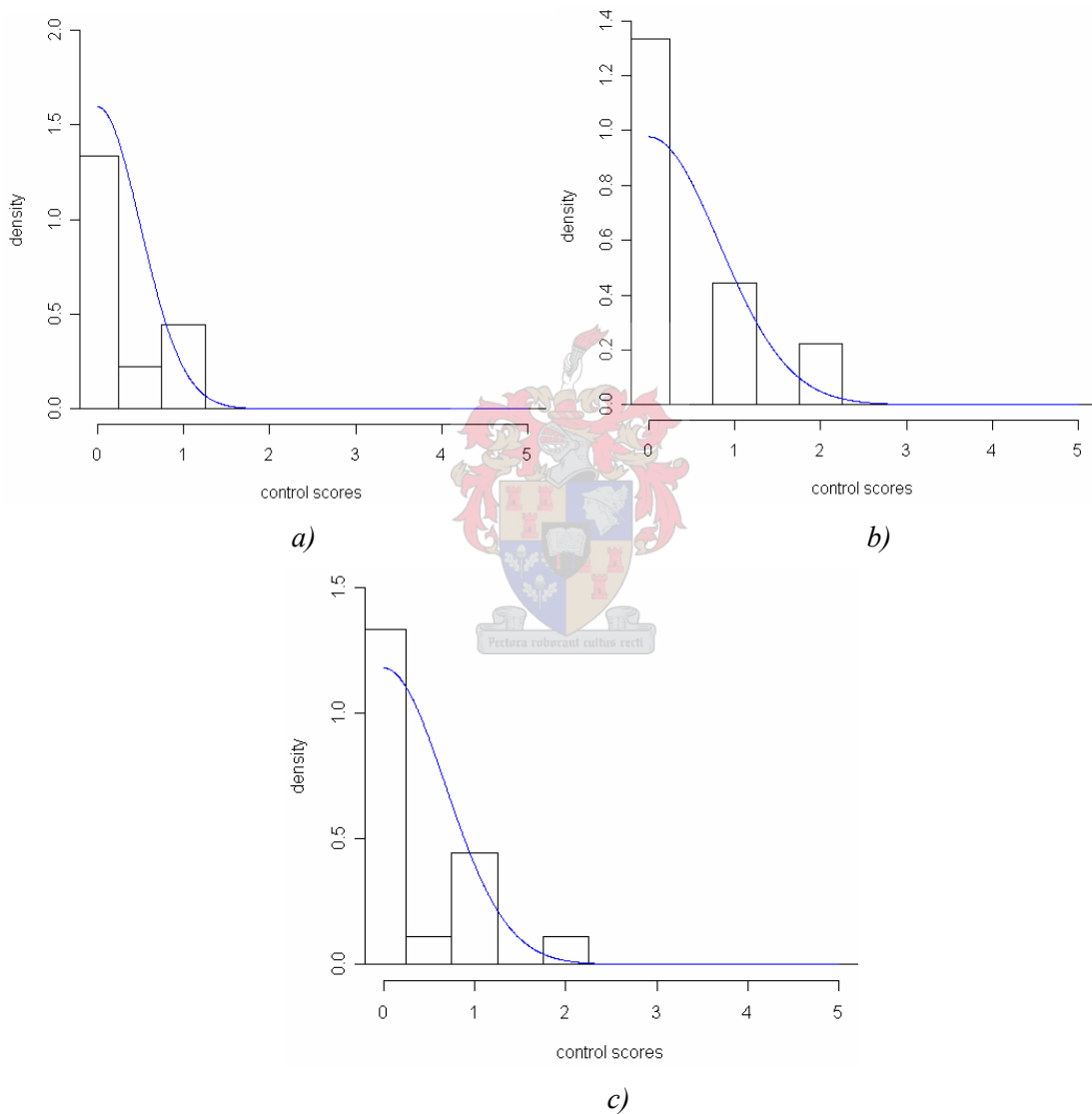


Figure 3.12: Histograms and corresponding estimated truncated absolute normal density function for the 3-day(a), 9-day(b) and combined control data, under the null hypothesis.

The significance levels² for the critical value of one prescribed by the IMQM test are 0.0455, 0.2207 and 0.1397, respectively for the control data of the 3-day, 9-day and

combined data set. Although the shape of the distributions are very similar, as displayed in Figures 3.13 a) – c), the obtained significance values corresponding to a critical value of one, differ substantially. Therefore the significance level used in the test seems to be larger than the typically used value of 0.05.

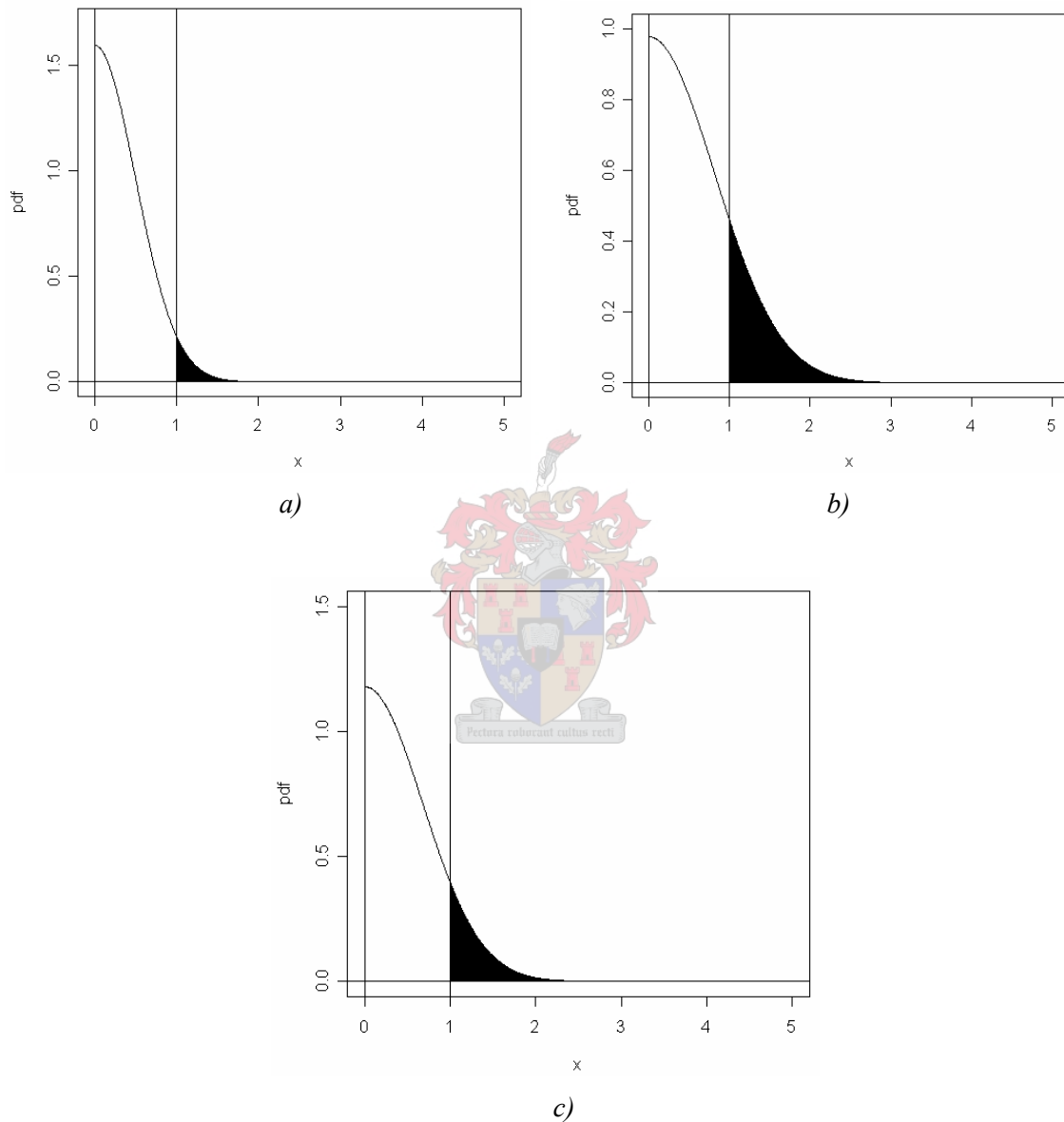


Figure 3.13: The density functions for the 3-day, 9-day and combined control data to illustrate the significance level for the critical value of one under the null hypothesis.

The observed significance levels for the means of 0.2778, 0.4444 and 0.3611 for the control scores of the appropriate data sets, are 0.5785, 0.5862 and 0.5937, respectively. Thus the null hypothesis that the mean of the coded control scores differs significantly

from zero cannot be rejected. This coincides with the results obtained in the IMQM procedure, but not with the t-test. Since it is known that the null hypothesis should be correct, the truncated absolute normal distribution thus renders acceptable results.

For the 3-day and 9-day data as well as the combined control data sets, the appropriate critical values to obtain a 5% significance level under the null hypothesis are estimated by once again considering equation (3.7). Critical values³ of 0.9800, 1.6002 and 1.3269 are obtained for the respective data sets. Note that these values differ substantially from those obtained for the test scores. Furthermore the estimated critical values vary somewhat between the different data sets although the samples are drawn from the same underlying distribution. This underlines the large influence of the estimated variance of the distributions. The estimated critical values for the 9-day control scores as well as the combined control scores are both larger than one (the critical value used in the MQM and IMQM procedures).

The estimates of the parameters¹ for the truncated distribution functions under the alternative hypothesis are (0.0020, 0.2500) for the 3-day control scores, (0.0066, 0.6667) for the 9-day control scores and (0.0055, 0.4583) for the combined score data. These estimates are used to calculate the probability of a type II error. These computed type II errors² under alternative distributions for the 3-day, 9-day and combined control scores are 0.9545, 0.7793 and 0.860, respectively. Although the estimates of μ are slightly larger than zero, the alternative distribution differs so little from the distribution under the null hypothesis that the type II errors are extremely large, resulting in low power.

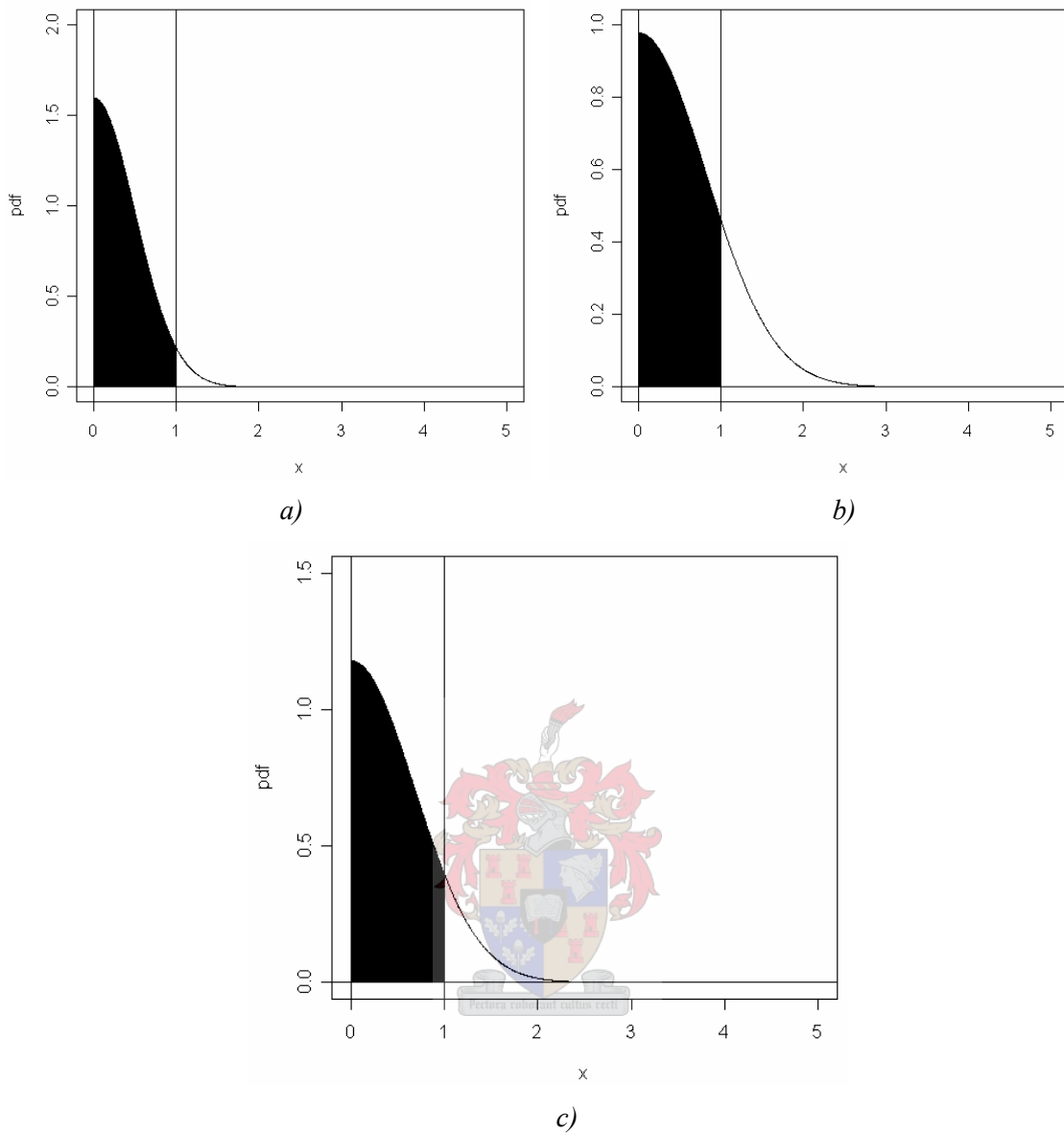


Figure 3.14: The density functions corresponding estimated truncated absolute normal distribution functions for the 3-day(a), 9-day(b) and combined control data, under the alternative hypothesis. The shaded area denotes the probability of committing a type II error.

When considering these results, the estimates for a critical value to use in the IMQM procedure are rather close to one which is currently used. The worrying factor however is that the distributions under the null and alternative hypothesis are so similar, that the probability of a type II error, i.e. not rejecting the null hypothesis when the alternative is true, is very large.

3.5 INFLUENCE OF PANEL SIZE

A feature of the MQM procedure that has not received much attention is that of the sample size. The panel is relatively small but it is of importance to examine what effect an even smaller sample size might have on the obtained results. This is necessary since the organisation under consideration deems a panel consisting of three judges together with repeating the trial a number of times to be permissible. Thus it is necessary to investigate whether the previous results would remain the same if the trial is completed a number of times with a smaller panel. This will be studied by firstly evaluating all possible combinations of size three to seven from the full panel and then determining how the results obtained by the procedure prescribed by the organisation would differ for these smaller “panels” compared to those of the full panel. The error rate is calculated as the proportion of samples that lead to a result that differs from that obtained using the complete panel. However, it is to be noted that this does not imply that the result obtained from the full panel is assumed to be correct.

Table 3.3 contains the error rates⁴ for the 3-day and 9-day trials, respectively, when panels of 3 to 7 are investigated. Keep in mind that the mean of the test data in the 3-day trial was 1.8889 and that of the 9-day trial was only 1.2222, leading to a significant result in both trials. Keep in mind however that the data in the 3-day trial are larger in magnitude and thus the mean of a sub-sample of them will typically be larger than that of a sub-sample from the 9-day test data. Since the full panel rejects the null hypothesis, a smaller panel not rejecting the null hypothesis, i.e. the mean of the panel not exceeding one, will be counted as an error. Therefore, since the mean of a sample from 3-day trial will typically be larger than that of a sample from the 9-day trial, the number of erroneous results will be less for the 3-day trial data. This explains why the error rates of the 3-day trial are much smaller than the corresponding error rates of the 9-day trial.

⁴ These error rates were obtained by implementing the R-function `repeat.taint.test()`, contained in the appendix, for the appropriate data set.

Table 3.3: Error rates for different panel sizes.

Panel size	3-day trial	9-day trial
3	0.1429	0.4524
4	0.0635	0.3730
5	0.0238	0.3095
6	0	0.2619
7	0	0.1667

A disadvantage of the previous investigation is that it compares the result obtained by a subset of the panel with that of the full panel, although it is unknown whether the result obtained by the full panel is correct. It would be useful to consider the error rate in a situation where it is known what the correct result should be. Therefore a similar study is carried out for the combined control data of the two trials, since it is known that no significant difference exists. In Table 3.4 these error rates are displayed and though they are much smaller it is still alarming since absolutely no difference actually exists.

Table 3.4: Error rates for the control data for different panel sizes.

Panel size	Control
3	0.0123
4	0.0033
5	0.0006
6	0.0001
7	0

To conclude...

Most of the test procedures in this chapter are parametric in nature and are therefore dependent on some distributional assumptions, thus the results will only be valid under these assumptions. In the following section the distribution of the test statistics used in previous sections will be under consideration to check whether the appropriate distributional assumptions are satisfied.

CHAPTER 4

BOOTSTRAP METHODOLOGY

4.1 BACKGROUND

Although parametric procedures are useful, they are hampered by the fact that they are only valid when their assumptions are satisfied. Moreover parametric solutions for a specific problem may often not be available or only asymptotically be available. This necessitates the implementation of non-parametric procedures. Advances in the processing power of computers have lead to the development of bootstrap methods that rely on Monte Carlo simulations instead of intricate mathematical derivations to obtain statistical results. These methods still apply the basic statistical principles and utilize probability theory to obtain needed measures of accuracy for statistical estimates and significance levels for inferential procedures.

Accuracy measures, such as standard errors, of estimates are vital for decision making, but for most estimates closed forms of such measures do not exist. The bootstrap procedure relies on the plug-in principle (Efron & Tibshirani, 1993) which estimates the unknown underlying distribution (F) from which the original sample was obtained by using resampling techniques and then estimates the standard error by the standard deviation of the distribution of the bootstrap replicates of the original sampling statistic. The bootstrap distribution is obtained by sampling with replacement from the empirical distribution associated with the original sample (of size n) in the non-parametric case or from the parametric distribution estimate in the parametric case. Parametric bootstrap distributions will not be implemented in this investigation of sensory data since it is known that assumptions concerning the underlying distribution of the data are not met. The probability distribution from which the sample \mathbf{x} is obtained can be presented as:

$$F \rightarrow \mathbf{x} = (x_1, x_2, \dots, x_n).$$

This sample may range from simple, for instance a vector or scalar, to complex, such as a matrix or a time-series. It is especially in these more complicated situations that the true value of the bootstrap is revealed. Figure 4.1 contains a schematic diagram similar to Efron & Tibshirani (1993, Figure 8.1), to illustrate how the bootstrap is implemented in the simple one sample case. The “Real World” represents the space containing the unknown underlying distribution from which the original sample was drawn. Thus the statistic under consideration ($\hat{\theta} = s(\mathbf{x})$) which may be very complex is estimated using the original sample. The distribution of this statistic is now approximated by drawing bootstrap samples \mathbf{x}^* with replacement from the original sample. The statistics are calculated for each bootstrap sample. These values will be referred to as the bootstrap replicates ($\hat{\theta}^* = s(\mathbf{x}^*)$) of the statistic. The bootstrap replicates are used to estimate the different properties of the statistic in question.

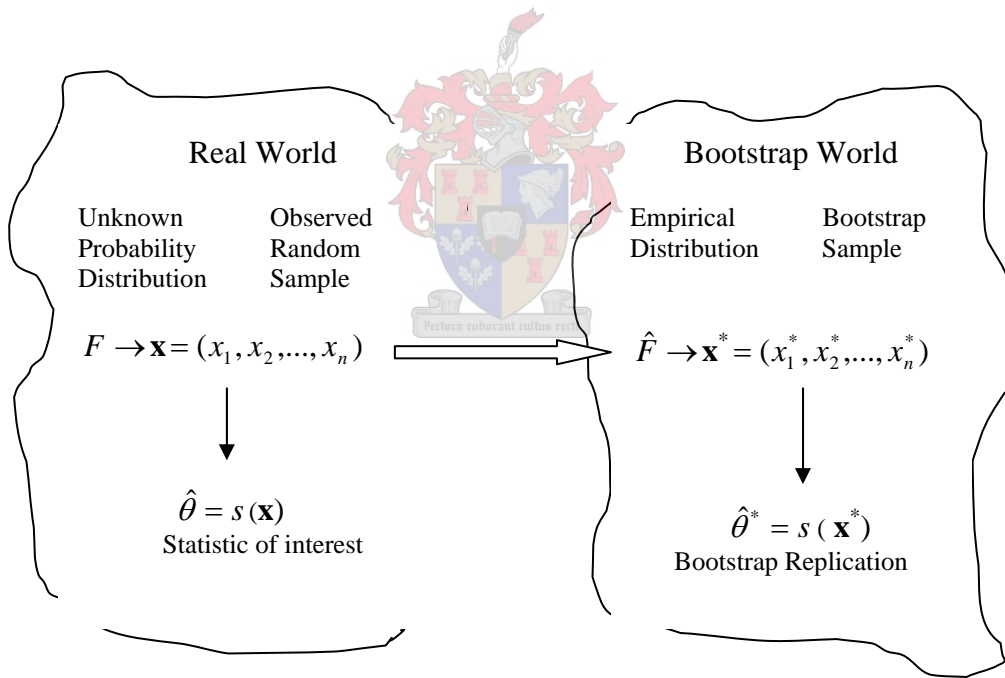


Figure 4.1: Schematic image used to illustrate the relationship between the original sample and the bootstrap replication.

4.2 APPLICATION⁵

4.2.1 Statistic in question

The aim of the sensory study considered in Chapter 3 is to verify whether there is a distinguishable difference between the control and treated pieces of chocolates considered in 3-day and 9-day trial. While this hypothesis was considered by implementing parametric testing methods in Chapter 3, non-parametric bootstrap methods will now be implemented. The implicit assumption that the control pieces are homogeneous is now under consideration. Since the control blocks are used as a reference in the t-test procedures, it is necessary to examine their properties further.

4.2.2 Bootstrap distribution of $\hat{\theta}$ for 3-day data

The t-test considered in Chapter 3 assumes that the differences between control chocolate pieces are distributed according to $N(\mu_c, \sigma_c^2)$ distribution and under the null hypothesis $\mu_c = 0$. A histogram approximating the bootstrap distribution of the mean of the control data in the 3-day data set is displayed in Figure 4.2. This was obtained by drawing 2000 bootstrap samples from the control data and computing the statistic in question which is the mean, for each of these. The dashed line indicates $\hat{\theta} = 0.2778$ which is the mean of the original control data. It can be seen that the shape of the distribution is similar to a normal distribution although not being completely symmetrical.

The t-test implemented in Section 3.2.1, determined that the mean does not differ significantly from zero. The t-test is however a parametric procedure and depends on the normality assumption that may not be valid. Thus this result is therefore only mentioned for reference purposes. Considering the histogram, it is clear that none of the bootstrap samples result in a mean of more than one; thus in none of these cases a significant result would have been obtained if the prescribed test used by the organisation was used.

⁵ Figures and estimates in this section is obtained by implementing R-function `bootstrap.dist()` included in the appendix.

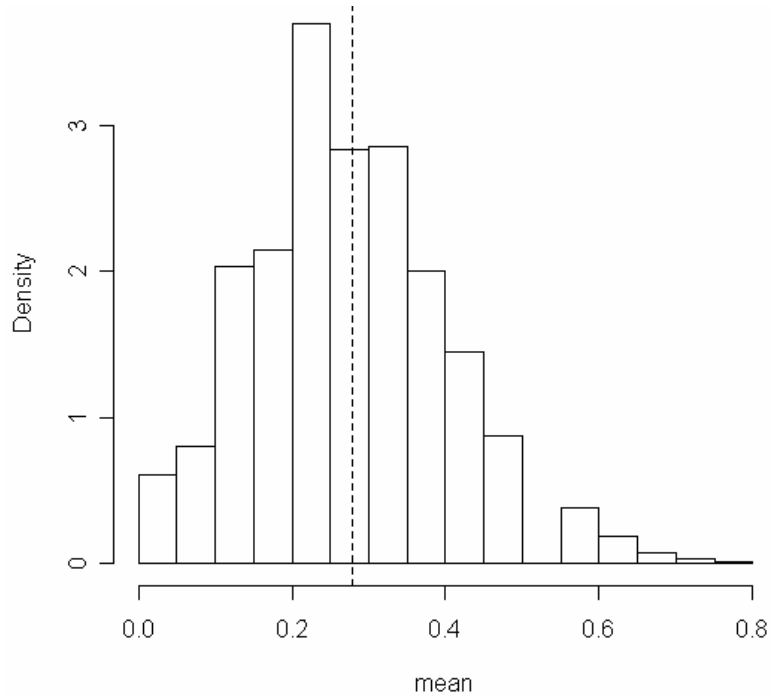


Figure 4.2: Bootstrap distribution of $\hat{\theta}$ for the control data of the 3-day data set.

4.2.3 Bootstrap distribution of $\hat{\theta}$ for 9-day data

Similarly the control set of the 9-day data is considered and the statistic under scrutiny is once again the mean. Using 2000 bootstrap replications the bootstrap distribution of the mean of the control scores is obtained as approximated by the histogram in Figure 4.3. The observed value for $\hat{\theta}$ is 0.4444, which is once again indicated by the dashed line. The t-test performed in Section 3.2.1, determined that the mean does not vary significantly from zero at a 5% significance level, but since this result relies on the normality assumption, it is once again only mentioned for the reference purposes.

The shape of the histogram does not seem to approximate a normal distribution since it contains gaps and thus does not seem to estimate a continuous distribution. It also does not seem to be symmetrical. This fact brings the compliance to the assumptions, on which the t-tests described in Section 3.2.1 are based, under suspicion.

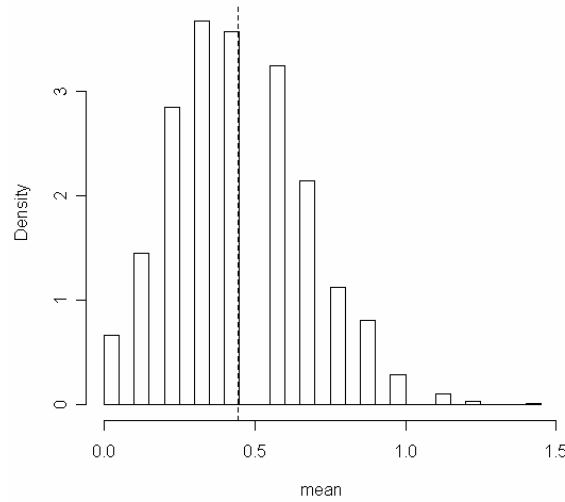


Figure 4.3: Bootstrap distribution of $\hat{\theta}$ for the control data of the 9-day data set.

Histograms are used to estimate the distributions under consideration. There are two options left to the discretion of the statistician when constructing a histogram, namely the bin width used as well as the bin origin. The bins are the non-overlapping intervals used to construct the histogram by counting the number of points in each bin (Scott, 1992). The bins are typically of the same width, called the bin width. Furthermore the bin origin refers to the position of the lower endpoint of a bin. Varying the bin width or the bin origin could lead to totally different histograms. Therefore the shape of a histogram may be manipulated by changes in the bin width and origin. Due to these shortcomings of histograms, other density estimates will also be applied to view the estimated underlying distribution. Since there were a restricted number of observed values allocated to the difference between the control blocks, only a finite number of possible values for the mean might exist which would explain the observed gaps in the distribution. This characteristic will be examined in Section 5.4.

4.2.4 Bootstrap distribution of $\hat{\theta}$ for the combined control data

It is assumed that the control pieces are distributed identically within each trial, as well as between these trials. Thus the estimated distribution of the mean of the control set of the two data sets combined will serve as $\hat{\theta}$ and should be similar to that obtained for the data sets separately. Differences between the above bootstrap distributions of the mean may be attributed to either differences between the panels used for the trials or random

noise which is incorporated in the control pieces used in these two trials. Random noise causes random variation within objects, but this noise is minimised by quality control procedures in manufacturing and is therefore assumed to be negligible.

The bootstrap replicates for the mean of the combined control data sets are computed by drawing bootstrap samples with replacement from the combined control data and calculating the mean for each of these. In the application of the bootstrap 2000 bootstrap replicates were used. The value of $\hat{\theta}$ is 0.3611 with an estimated standard error of 0.1389, which once again implies that $\hat{\theta}$ differs significantly from 0 at a 5% significance level since the associated significance level is 0.0093 when using a t-distribution with 17 degrees of freedom as in Section 3.2.1. Here this parametric result is once again included for reference purposes sake.

The histogram of the bootstrap replicates in Figure 4.4, which approximates the distribution of $\hat{\theta}$ for the combined control set described above, does not have similar features to Figure 4.3, but appears similar to that of Figure 4.2. Further examination of the distributions is necessary to come to any definite conclusions concerning this distribution.

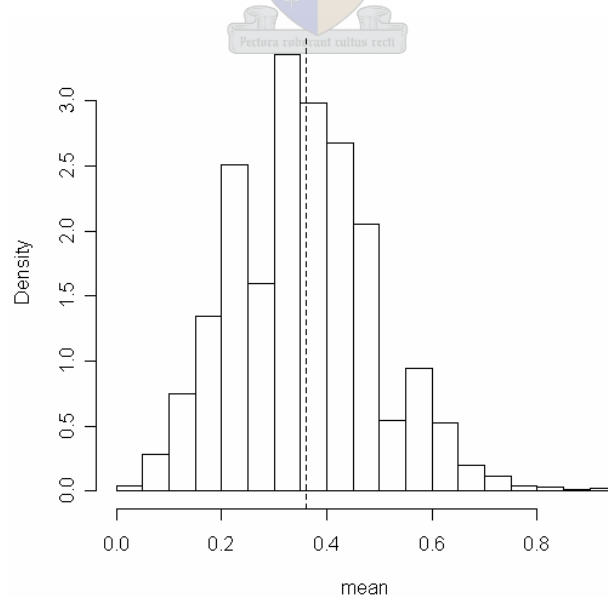


Figure 4.4: Bootstrap distribution of $\hat{\theta}$ for the control data of combined data set.

4.2.5 Density estimates of Bootstrap distribution of $\hat{\theta}$ ⁶

Although histograms are convenient and straightforward to implement there are disadvantages associated with them. The fact that there is no unanimous optimal bin width together with a choice of the bin origin that is left largely to the discretion of the statistician, lead to different diagrams for the same data set. These choices of bin width and starting point could be very influential and could therefore manipulate or obscure the true properties of the distribution. A shifted average histogram is proposed by Scott (1992) to address this issue.

Another technique for obtaining a density estimate is to use a smoothing function, such as cubic smoothing spline functions, to estimate the underlying distribution. These functions typically are of the form:

$$J_{\lambda}(f) = \sum_1^n [y_i - f(z_i)]^2 + \lambda \int [f''(z_i)]^2 dz \quad (4.1)$$

where: $f(\cdot)$ is a continuous function that is at least two times differentiable,

λ is an adjustable smoothing parameter,

z_i is an independent variable and y_i is the dependent variable.

This function is fitted to the observed data points as independent variables and the empirical distribution as the dependent variable. The first term in equation (4.1) is a term which indicates the error made by estimating the dependent variable with $f(z_i)$ and the second term is a penalty term that adds a penalty for the irregularity of the fitted curve. Different values of the smoothing parameter will lead to different optimal curves (but at the expense of an increase in the bias). The larger the smoothing parameter, the smoother the obtained curve. Similarly, small values of the λ will lead to jagged curves that follow the data points rigidly, but with an increase in the variance.

Figure 4.5 contains an illustration of the curve obtained by applying a specific variation of a cubic smoothing spline to bootstrap distributions of $\hat{\theta}$ for the 3-day control data, obtained in Section 4.2.2. A kernel smoothing method using Gaussian kernel smoothing

⁶ Figures in this section are obtained by implementing R-function `bootstrap.dens()` included in the appendix

was implemented. The red line is drawn to underline that a mean value of less than zero is inadmissible, since the judges were not allowed to award negative scores. The shape of the solid curve does not differ noticeably from that of a normal density function, as implied by the histogram. One important difference is the rather large portion of the estimated normal distribution that falls outside the admissible region demarcated by the red line.

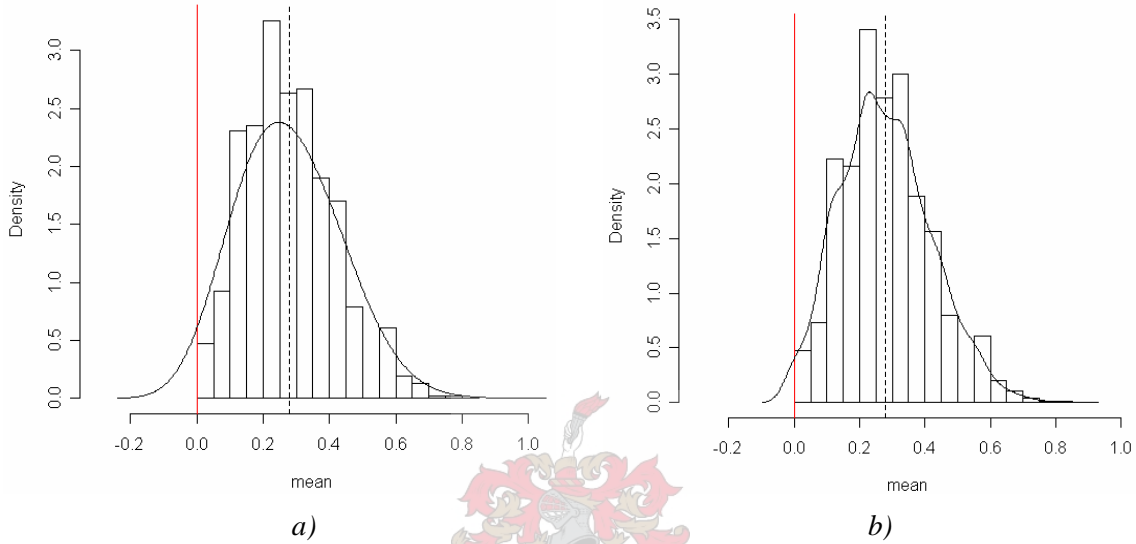


Figure 4.5: Solid curve obtained by applying cubic smoothing splines superimposed on histogram of bootstrap distribution of $\hat{\theta}$ for the control data of 3-day data for different values of λ . The red line demarcates the admissible values of the mean of the mean, while the dashed line indicates the original estimate $\hat{\theta}$.

This procedure was also applied to the bootstrap replicates obtained in Section 4.2.2 for the control data of the 9-day data set which as previously stated did not seem to comply with the assumption of having a continuous distribution. Figure 4.6 contains estimates of a continuous density function using different values of the smoothing parameter. It is apparent in (a) that a smooth normal function does not fit the data well but on the other hand, the curves in (c) and (d) are very ragged and do not seem to be of the form of a normal density function.

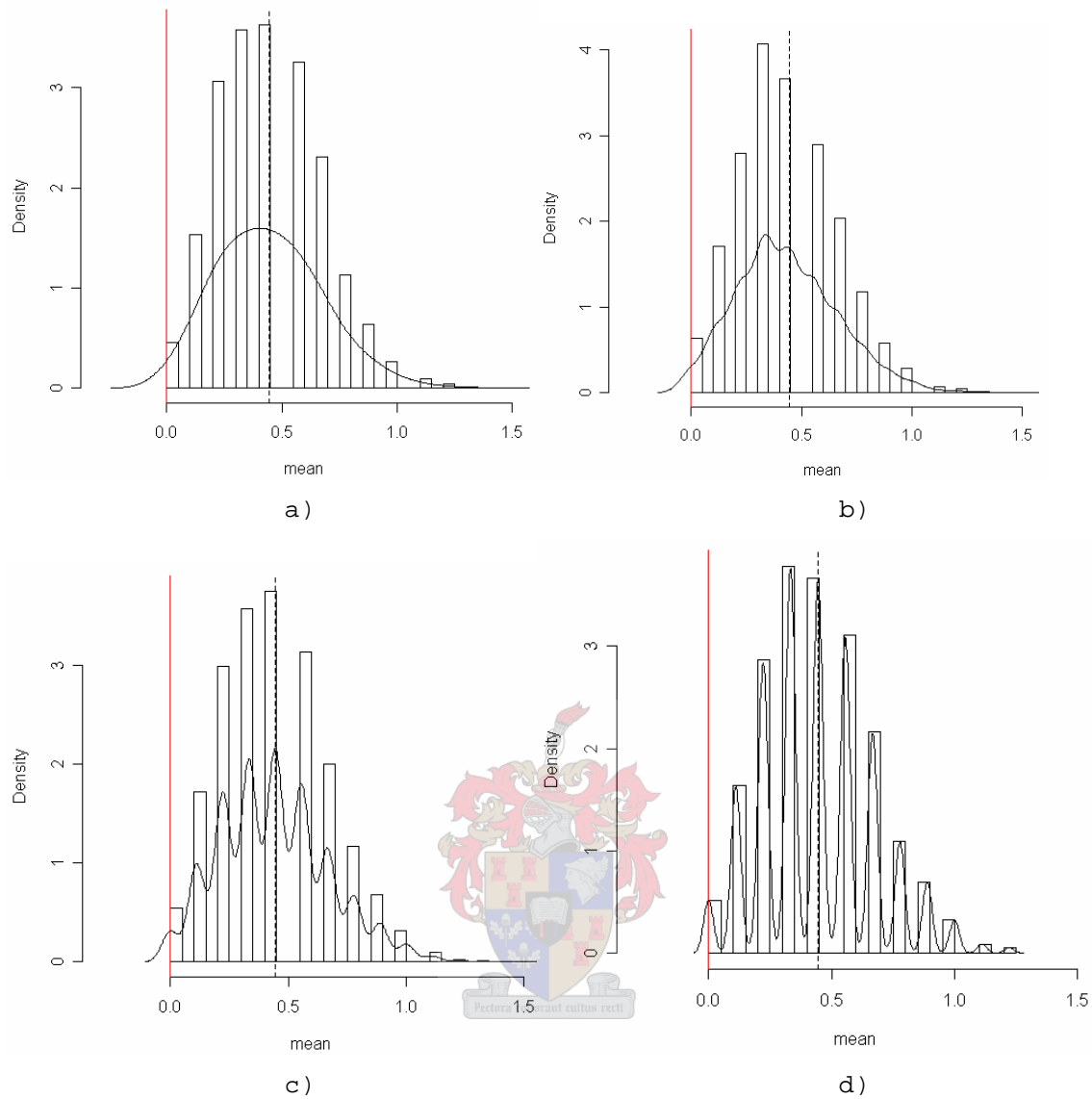


Figure 4.6: Solid curve obtained applying cubic smoothing spline superimposed on histogram of Bootstrap distribution of $\hat{\theta}$ for the control data of 9-day data for different values of λ . The red line demarcates the admissible values of the mean of the mean, while the dashed line indicates the original estimate $\hat{\theta}$.

Figure 4.7 displays the solid curve superimposed on the histogram obtained for the bootstrap distribution of $\hat{\theta}$ for the control data of the combined data. This curve, as expected, appears to have the same form as a normal density curve although negative values are once again not permitted.

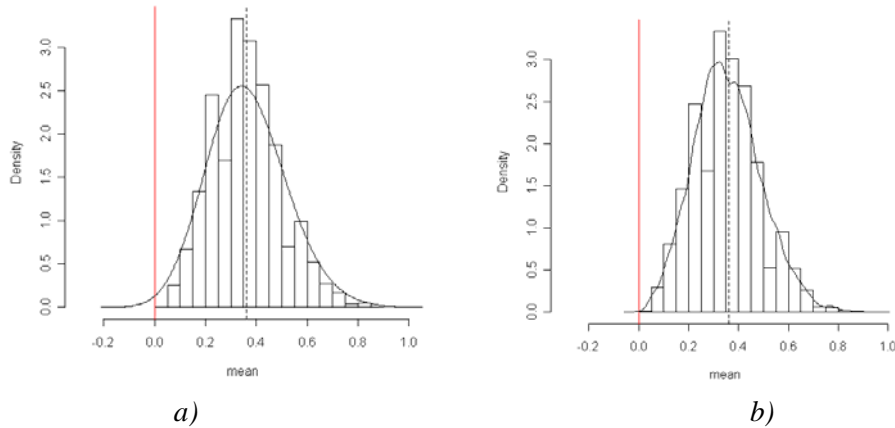


Figure 4.7: Illustration of the smooth curve superimposed on the histogram of the Bootstrap distribution of $\hat{\theta}$ for the control data of combined data set for different values of λ . The red line demarcates the admissible values of the mean of the mean, while the dashed line indicates the original estimate $\hat{\theta}$.

4.3 BOOTSTRAP TESTS⁷

The bootstrap methodology may also be implemented to test a specific hypothesis. As discussed in some of the previous sections, the t-test would typically be implemented to test for a significant difference between two sample means, but the organisation that carried out this trial has its own standard test for determining whether a significant difference exists. The bootstrap test will now be used as a non-parametric procedure to compare these methods.

The one-sided bootstrap test consists of evaluating a specified test statistic for a large number of bootstrap samples and subsequently calculating the proportion of the bootstrap replicates that exceed the observed test statistic. Efron & Tibshirani (1993) refer to this proportion as the achieved significance level (ASL). The ASL is compared to the required significance level, and if the ASL is smaller the null hypothesis is rejected.

Thus for the paired one-sample t-test the appropriate test statistic is :

$$\hat{\theta} = t(\mathbf{y}) = t(y_1, \dots, y_n) = \frac{\bar{y}}{s\hat{e}(\bar{y})}, \quad \text{where } \mathbf{y} \text{ is defined as in Section 3.4.2.}$$

⁷ The bootstrap tests were performed by applying the R-function `boot.test()`, contained in the appendix, to the appropriate data set. If only one data vector is considered the R-function `one.smpl.boot.test()` is used.

The corresponding two-sample test statistic is:

$$\hat{\theta}' = t(\mathbf{y}) = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\bar{\sigma}_T^2/n + \bar{\sigma}_C^2/m}}, \quad \text{where } \bar{\sigma}_T^2 = \sum_i^n (x_{Ti} - \bar{x}_T)^2 / (n-1)$$

$$\text{and } \bar{\sigma}_C^2 = \sum_i^m (x_{Ci} - \bar{x}_C)^2 / (m-1).$$

The test statistic for the standard test is defined as:

$$\hat{\theta}_{MQM} = t(\mathbf{y}) = t(x_1, \dots, x_n) = \bar{x}_T, \quad \text{where } \bar{x}_T \text{ represents the mean of}$$

the test scores.

The bootstrap replications of \mathbf{y} consist of drawing samples of size n with replacement from the original \mathbf{y} vector and then calculating the appropriate test statistic. This is called a one-sample t-test since the \mathbf{y} vector is sampled and not the \mathbf{x}_T and \mathbf{x}_C vector separately. This is done since the difference between the test score and control score awarded by a specific individual is under consideration and if sampling was to be done separately this dependency would be lost. To explore the effect of ignoring this dependency, the two sample methodology will also be carried out for the t-test and the results compared to the one-sample procedure. The two-sample procedure consists of pooling the test and control scores, drawing $2n$ observations with replacement and then assigning the first n to the test vector and the rest to the control. Since there is no reason to assume that the variation of the test and control items are the same, the test statistic being evaluated is $\hat{\theta}'$, defined above. For the standard procedure applied by the organisation the bootstrap sample consists of drawing a sample of size n with replacement from the original test scores and simply calculating the mean.

Tables 4.1, 4.2 and 4.3 contain the results obtained by using the bootstrap testing methodology described above, for each data set as well as the p-value obtained for the corresponding one-sample as well as two-sample t-test. It is apparent that the results obtained by the organisation's prescribed test do not correspond to those results found in either of the parametric t-test or the non-parametric bootstrap applications. The results of the one- and two-sample t-tests do not differ significantly, which could imply that the test and control scores are awarded independently.

Table 4.1: Results for the bootstrap test for the different test procedures for the 3-day trial.

Trial	Testing procedure	Null Hypothesis	$\hat{\theta}$	Parametric t-test: Significance level	Bootstrap test: ASL	Conclusion based on ASL
3-day trial	One sample t-test	$H_0 : \mu_{C3} = 0$	1.8898	0.0477	0.0200	Reject the null hypothesis at 5% significance level
	One sample t-test	$H_{01} : \mu_{T3} = \mu_{C3}$	3.4296	0.0045	0.0050	Reject the null hypothesis at 5% significance level
	Two sample t-test	$H_0 : \mu_{T3} = \mu_{C3}$	3.3712	0.0019	0.0025	Reject the null hypothesis at 5% significance level
	One sample t-test	$H_0 : \mu_{T3} = 0$	4.1538	0.0016	0.0030	Cannot reject the null hypothesis at 5% significance level
	IMQM test for 1	$H_0 : \mu_{T3} = 0$	1.8889	unknown	0.9845	Cannot reject the null hypothesis at 5% significance level
	IMQM test for mean	$H_0 : \mu_{T3} = 0$	1.8889	unknown	0.5440	Cannot reject the null hypothesis at 5% significance level
	IMQM test for 2.5556	$H_0 : \mu_{T3} = 0$	1.8889	unknown	0.0805	Cannot reject the null hypothesis at 5% significance level

Table 4.2: Results for the bootstrap test for the different test procedures for the 9-day trial.

Trial	Testing procedure	Null Hypothesis	$\hat{\theta}$	Parametric t-test: Significance level	Bootstrap test: ASL	Conclusion based on ASL
9-day trial	One sample t-test	$H_0 : \mu_{C9} = 0$	1.8353	0.0519	0.0210	Reject the null hypothesis at 5% significance level
	One sample t-test	$H_{04} : \mu_{T9} = \mu_{C9}$	2.4010	0.0216	0.0295	Reject the null hypothesis at 5% significance level
	Two sample t-test	$H_0 : \mu_{T9} = \mu_{C9}$	2.1106	0.0255	0.0310	Reject the null hypothesis at 5% significance level
	One sample t-test	$H_0 : \mu_{T9} = 0$	4.4000	0.0011	0.0050	Cannot reject the null hypothesis at 5% significance level
	IMQM test for 1	$H_0 : \mu_{T9} = 0$	1.2222	unknown	0.8610	Cannot reject the null hypothesis at 5% significance level
	IMQM test for mean	$H_0 : \mu_{T9} = 0$	1.2222	unknown	0.5910	Cannot reject the null hypothesis at 5% significance level
	IMQM test for 1.6667	$H_0 : \mu_{T9} = 0$	1.2222	unknown	0.0565	Cannot reject the null hypothesis at 5% significance level

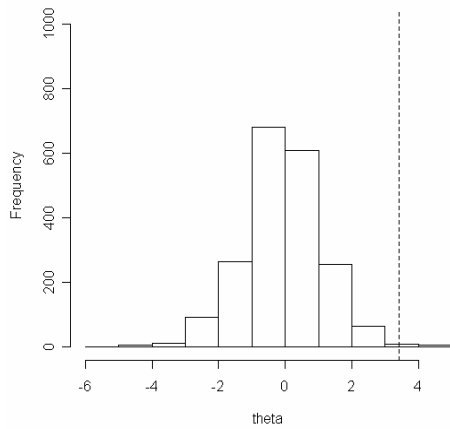
Table 4.3: Results for the bootstrap test for the different test procedures for the combined data sets.

Trial	Testing procedure	Null Hypothesis	$\hat{\theta}$	Parametric t-test: Significance level	Bootstrap test: ASL	Conclusion based on ASL
Combined data	Two sample t-test	$H_{02} : \mu_{T9} = \mu_{C3}$	3.0052	0.00419	0.0045	Reject the null hypothesis at 5% significance level
	Two sample t-test	$H_{03} : \mu_{C9} = \mu_{C3}$	0.5883	0.2823	0.3103	Cannot reject the null hypothesis at 5% significance level
	Two sample t-test	$H_{05} : \mu_{T3} = \mu_{C9}$	2.8037	0.0064	0.0065	Reject the null hypothesis at 5% significance level
	Two sample t-test	$H_{06} : \mu_{T9} = \mu_{T3}$	-1.2511	0.8856	0.8955	Cannot reject the null hypothesis at 5% significance level
	One sample t-test	$H_0 : \mu_C = 0$	2.6000	0.0093	0.0010	Reject the null hypothesis at 5% significance level

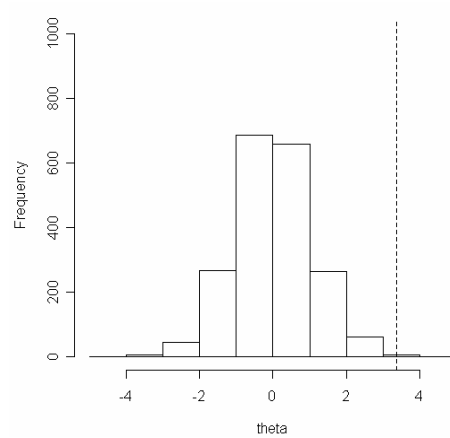
In Chapter 3 candidates for possible critical values were estimated. When these estimated critical values were used when implementing the bootstrap replicates of the IMQM tests, the ASL approaches zero. One possible reason for this is the inadequacy of the continuous truncated absolute normal distribution in estimating the true underlying distribution. To estimate the appropriate critical value for a specified 5% significance level the 95th percentile of the bootstrap distribution for the mean was determined. The corresponding ASL-values are shown in Tables 4.1 and 4.2. For the 3-day trial the optimal critical value obtained is 2.5556 and the corresponding value for the 9-day trial is 1.6667. The corresponding ASL-values are slightly larger as the specified 5% due to replication of bootstrap mean values. When comparing these ASL-values with those associated with a critical value of one the inadequacy of the latter criterion is clearly demonstrated.

These critical values are to be used in the organisation's prescribed test procedure instead of the current value of one. The variation between these two critical values is large, which underlines the criticism cited earlier concerning the fact that this test procedure does not take the variation within the data into account. For instance, a panel of ten may award the following scores (1, 1, 1, 1.5, 1, 0.5, 1, 1, 1.5, 0.5), while another panel might award scores of (0.5, 0, 0, 0, 0, 0, 5, 0, 4, 0.5) which would both lead to a mean of one. In the first case all the panellists perceived a difference while in the second panel only two of panellists were able to pick up any difference. The result from the first panel seems more reliable since they consistently experienced a difference while in the second panel this was not the case. When implementing the prescribed MQM procedure one would obtain the same result for both panels, which is disturbing.

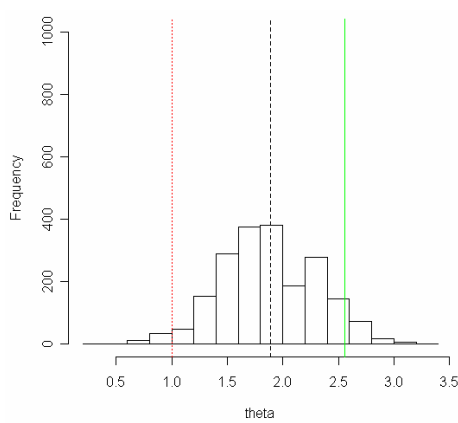
Figure 4.8 contains the histograms illustrating the bootstrap distributions of the various test statistics. The red line in panels c) and e) indicate the IMQM critical value of one and the green line the criterion associated with an approximately 5% significance level. None of the histograms seem to be estimating a symmetrical distribution.



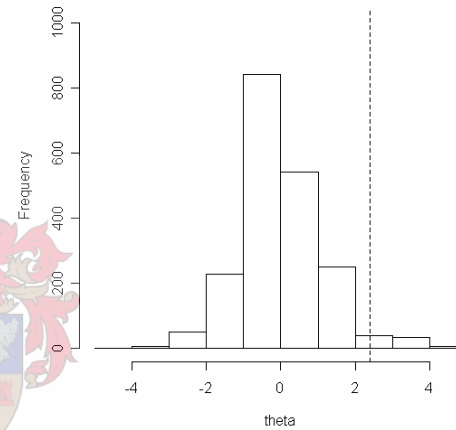
a) Histogram of one-sample t -test for 3-day data.



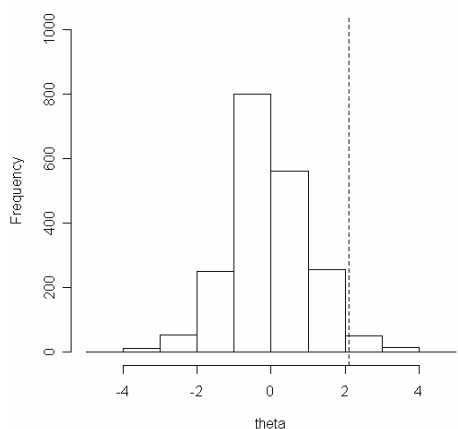
b) Histogram of two-sample t -test for 3-day data.



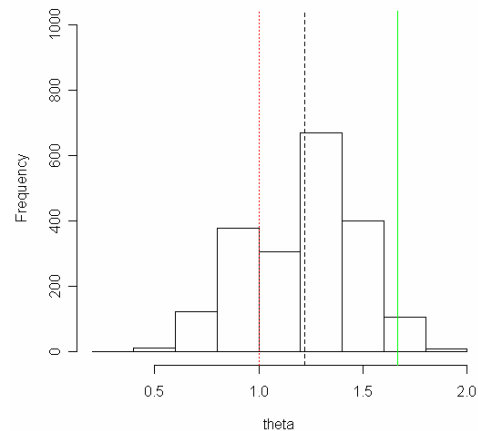
c) Histogram of standard procedure for 3-day data.



d) Histogram of one-sample t -test for 9-day data.



e) Histogram of one-sample t -test for 9-day data.



f) Histogram of standard procedure for 9-day data.

Fig.4.8: Histograms approximating the distributions of the test statistics for the one-sample and two-sample test procedures for either data set.

The distributions of the appropriate test statistics have been explored in some detail as well as the accuracy of the obtained tests. Due to the lack of symmetry of the bootstrap distributions the mean may not be an appropriate test statistic. In the following section the choice of mean as test statistic is under scrutiny.

4.4 APPLYING PCA-BIPLOTS

The suitability of the choice of the mean as test statistic is now of interest. The mean of the control data is subtracted from the mean of the test data, and this difference served as the test statistic. Whether this is the optimal choice for discriminating between objects is important. Since the shape of the histogram of the test scores displayed in Section 3.3 deviates considerably from being symmetrical, it is expected that the mean may be an inadequate test statistic. Several candidates for a test statistic do exist however. A statistic such as the median may also be appropriate due to the shape of the data or maybe a combined measure of all deciles might be more representative of the distribution. All deciles as well as the mean will be considered as candidates for a test statistic.

Bootstrap samples will now be used to gain some insight into the possible distribution of some of these possible test statistics. Since histograms will not be able to illustrate the different relationships between the above candidate statistics, biplots will be implemented for this objective. The biplot methodology is discussed in some detail in Chapter 6 and will now simply be implemented as an exploratory tool. All the test statistics will be calculated for each bootstrap sample. These candidate test statistics will serve as our variables and the set of bootstrap samples for the control and test data will serve as our sample. Histograms of each of the candidate test statistics for our sample will be displayed separately.

The PCA-biplot⁴ in Figure 4.9 represents 200 bootstrap samples from the 3-day data. For each of these the 10th, 20th, 30th, 40th, 50th (median), 60th, 70th, 80th and 90th percentiles as well as the mean of the difference between the test and control score were

⁴ The R-function `boot.test.bipl()` contained in the appendix was used in conjunction with the functions `PCA.bipl()`, `blegend()` and `drawbipl()` utilising the R-library **MASS**.

evaluated. Thus each bootstrap sample is regarded as an observation with these statistics as the variables. Each bootstrap replicate is represented by a point on the biplot and each statistic by an axis. Most of the samples seem to cluster around the mean axis or form lines parallel to the mean. An interesting feature here is that the angle between the axes representing the mean and the median is larger than the angle between the mean and the 60th percentile. This implies that the 60th percentile and the mean are more correlated. This may be attributed to the lack of symmetry observed in the distribution of the 3-day data set. The 50th percentile is the best representative however, since it lies almost parallel to first eigenvector which forms the horizontal axis.

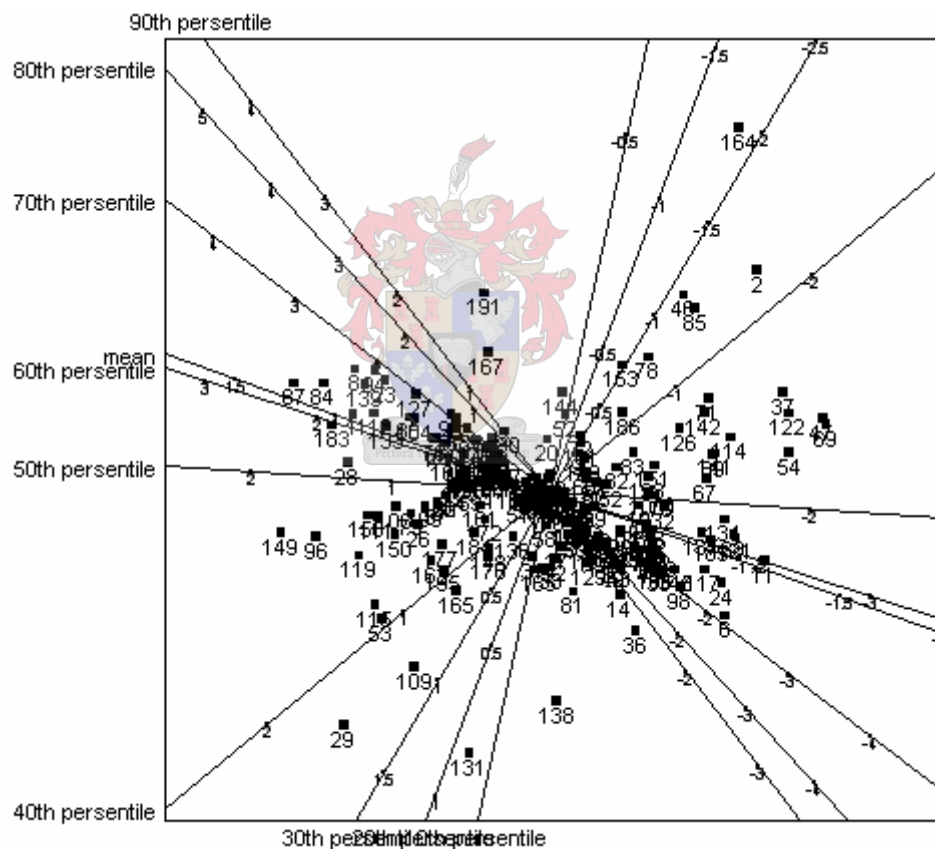
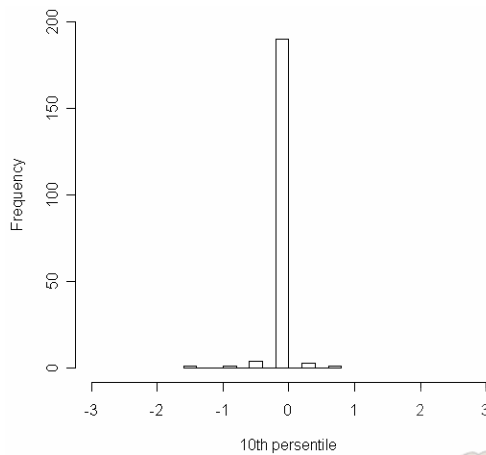


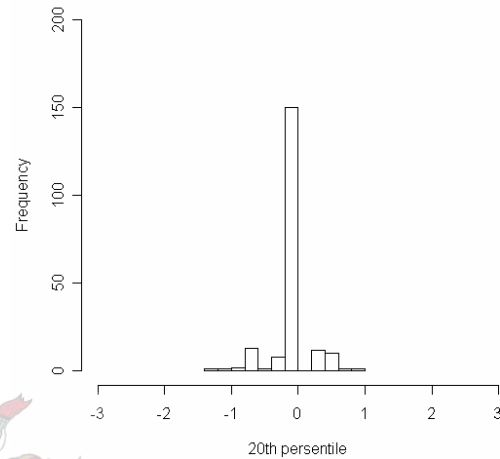
Figure 4.9: Biplot of candidates for test statistics for 200 bootstrap samples from the 3-day data.

Figure 4.10 a) to j) contain the histograms estimating the bootstrap distributions for the respective candidate test statistics. For the lower percentiles very few different values were encountered. The bootstrap distribution of the mid-range percentiles appear to

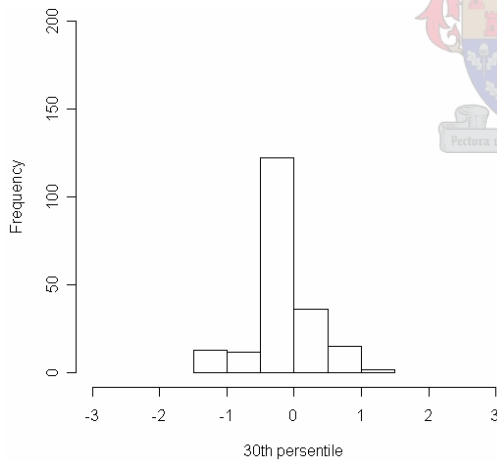
become more normal in shape. When considering the Figures 4.10 d) and e), the difference in shape is prominent. The mean does however seem to approximate a normal distribution more closely, and thus may be a more appropriate choice for a test statistic.



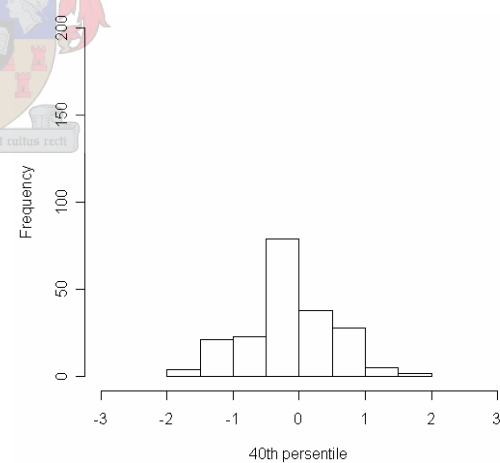
a) Histogram of the 10th percentile for 200 bootstrap samples.



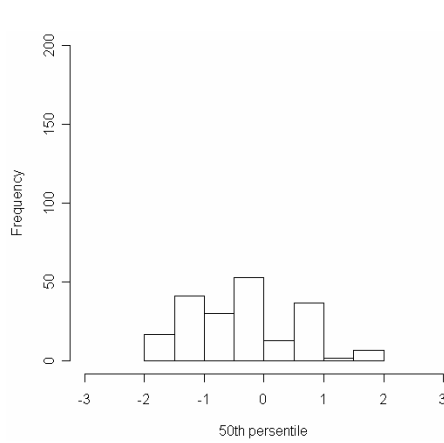
b) Histogram of the 20th percentile for 200 bootstrap samples.



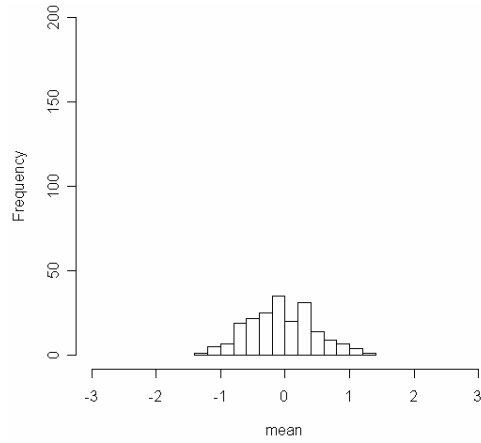
c) Histogram of the 30th percentile for 200 bootstrap samples.



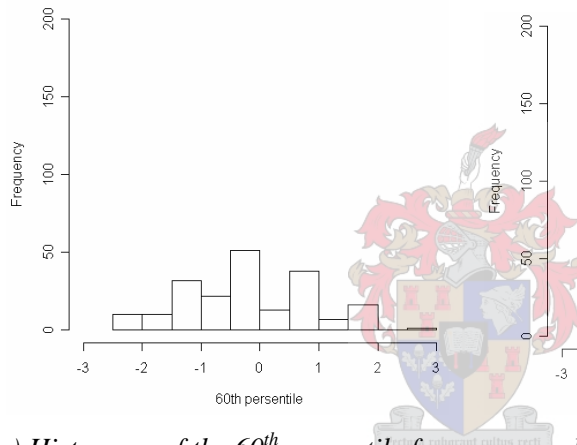
d) Histogram of the 40th percentile for 200 bootstrap samples.



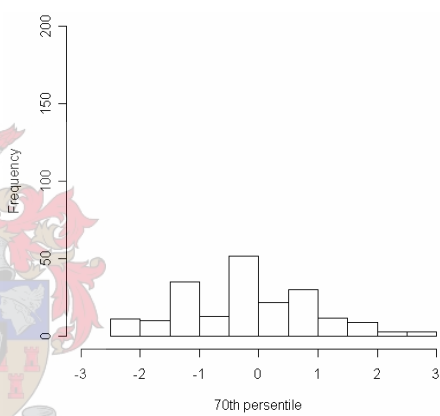
e) Histogram of the 50th percentile for
200 bootstrap samples.



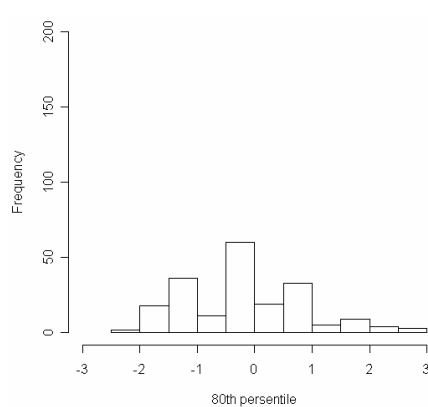
f) Histogram of the mean for 200 for
200 bootstrap samples.



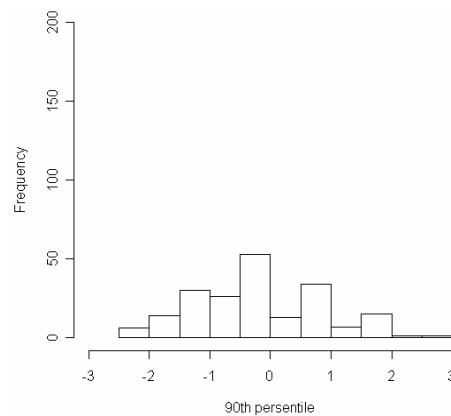
g) Histogram of the 60th percentile for
200 bootstrap samples.



h) Histogram of the 70th percentile for
200 bootstrap samples.



i) Histogram of the 80th percentile for
200 bootstrap samples.



j) Histogram of the 90th percentile for
200 bootstrap samples.

Figure 4.10: Histograms of the candidate test statistics for 200 bootstrap samples from the 3-day data.

Figure 4.11 contains the biplot for the 9-day data similar to that in Figure 4.9. Now the axis representing the median does correspond quite closely to that of the mean. However the median seems to be a better choice for a test statistic, since it lies in the same direction as the first eigenvector and thus explains the most variation within the data. The samples seem to cluster around the mean once again. Obvious parallel clusters are not observed. Figure 4.12 e) and f) estimate the bootstrap median and mean, respectively. Although their range is similar, the obvious difference between these two statistics is clearly illustrated, namely that one is continuous and the other discrete. Once again the histograms of the lower percentiles show that only a few distinct values for these statistics were encountered.

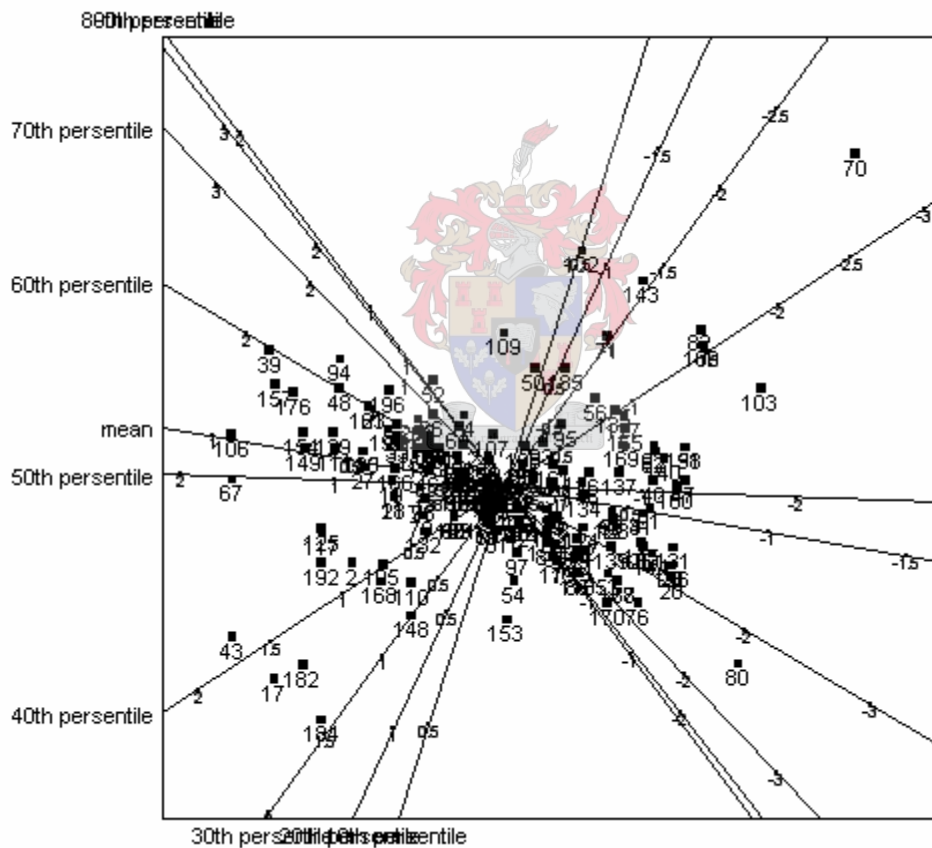
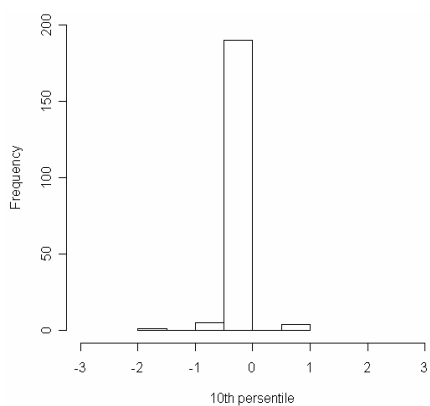
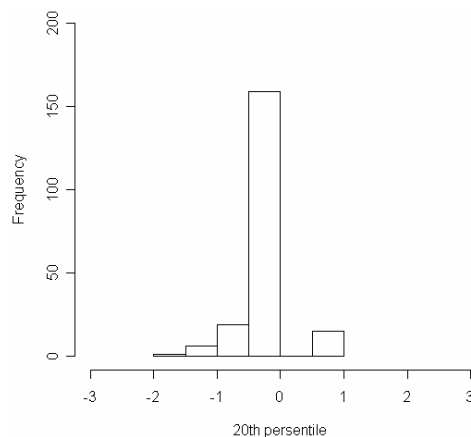


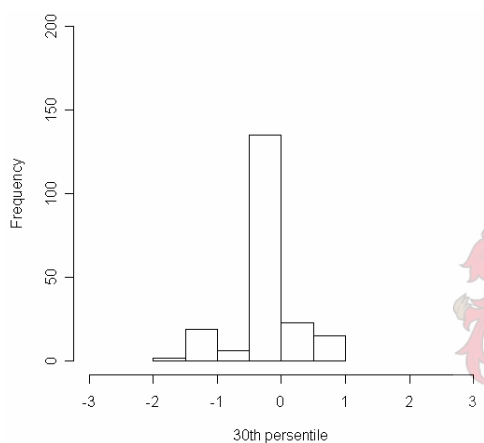
Figure 4.11: Biplot of candidates for test statistics for 200 bootstrap samples from the 9-day data.



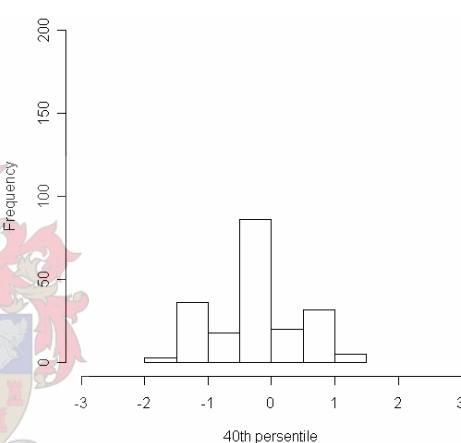
a) Histogram of the 10th percentile for 200 bootstrap samples.



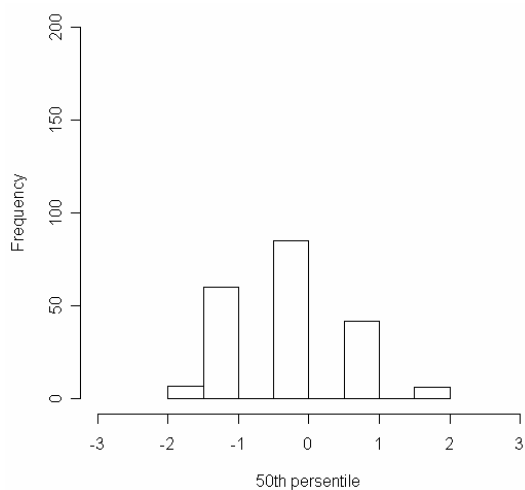
b) Histogram of the 20th percentile for 200 bootstrap samples.



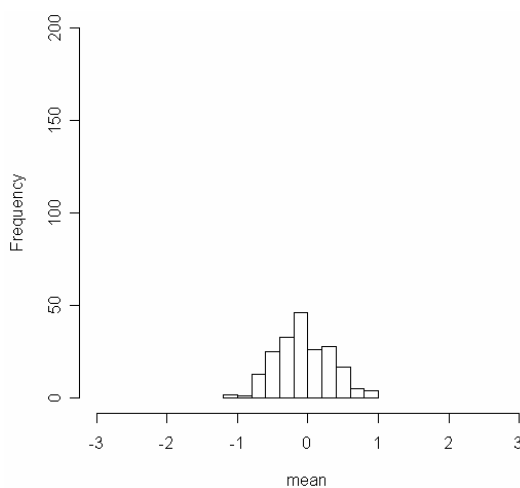
c) Histogram of the 30th percentile for 200 bootstrap samples.



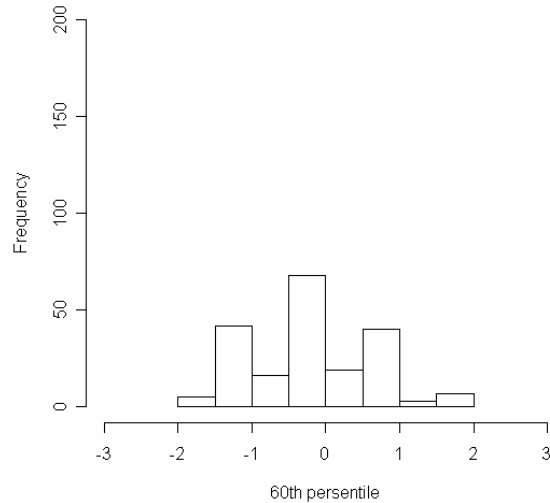
d) Histogram of the 40th percentile for 200 bootstrap samples.



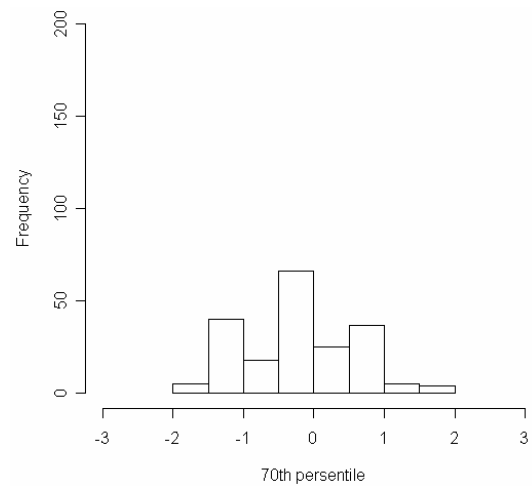
e) Histogram of the 50th percentile for 200 bootstrap samples.



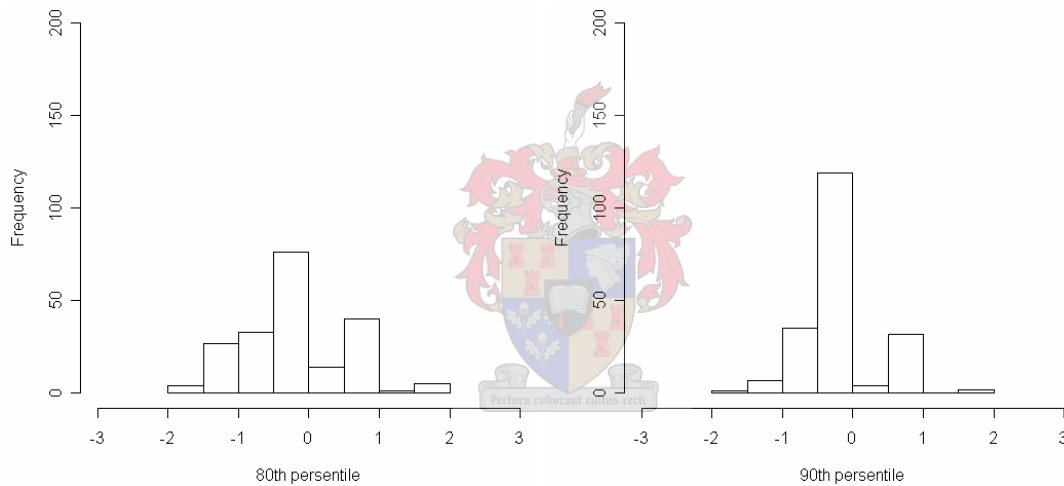
f) Histogram of the mean for 200 bootstrap samples.



g) Histogram of the 60th percentile for 200 bootstrap samples.



h) Histogram of the 70th percentile for 200 bootstrap samples.



i) Histogram of the 80th percentile for 200 bootstrap samples.

j) Histogram of the 90th percentile for 200 bootstrap samples.

Figure 4.12: Histograms of the candidate test statistics for 200 bootstrap samples from the 9-day data.

The size vector is a combination of the deciles that represent the variation in the data as well as the distribution of the data. From the biplots in Figure 4.11 and 4.12 it is clear that the mean will be an adequate size vector since both axes representing the mean are parallel to the respective horizontal axis.

To conclude...

Some knowledge of the distribution of certain test statistics as well as the accuracy of typical parametric as well as IMQM tests have been obtained in this chapter. Non-parametric bootstrap tests were also implemented to test the appropriate hypotheses. In the following chapter other non-parametric test procedures will be implemented viz. permutation tests. These will be used to gain further perspective on the accuracy of the previous results.



CHAPTER 5

PERMUTATION TESTS

5.1 BACKGROUND

The original reason for the development of permutation tests was to support R.A. Fisher's arguments for suggesting the Student's t-distribution as an appropriate distribution used in hypothesis testing (Efron & Tibshirani, 1993). Due to the vast improvements in computer power and processing speed this procedure has now become a powerful, easy to implement and widely applicable tool in the testing of a variety of hypotheses.

Permutation testing is a non-parametric procedure and thus may give more reliable results than those obtained from test methodologies relying on the assumptions that are not completely met by the data under consideration. The main advantage is that very few assumptions need to be satisfied in order to apply the permutation test procedure and thus for most parametric or non-parametric tests a permutation test equivalent exists. The main assumption that has to be met is that under the null hypothesis the distribution of the two samples under consideration must be the same (Efron & Tibshirani, 1993). This is called a symmetrical hypothesis, and is a very weak assumption.

In Chapter 4 the bootstrap methodology was utilized as a non-parametric procedure to test for differences between the two samples. Two types of hypotheses were considered: two-sample problems where the distributions of the two samples were compared with each other and one-sample problems where a specific characteristic, for instance the mean, of the distribution was considered. In the former of these two the assumption under the null hypothesis is that the two distributions are identical. This hypothesis can also be investigated by making use of the permutation test methodology, since the assumption of symmetry, described earlier, is met.

In the one-sample case the null hypothesis only makes assumptions about the distribution of one sample, thus the required symmetry does not exist and the one-sample problem cannot be addressed by a permutation test. When two sets of scores are deemed to be related and for instance the difference between them is used as the statistic considered, the two-sample problem reduces to a one-sample problem since a specific aspect of these differences is then under consideration. Thus a permutation test cannot be employed to consider paired comparison problems. In order to implement a permutation test for paired scores, the dependence between a pair of scores is ignored and thus the two-sample procedure is used.

The logic underlying the permutation test can be described as follows (cf. Efron & Tibshirani, 1993): Consider two samples \mathbf{z} and \mathbf{y} drawn from distributions F and G , respectively. The question being considered is whether these two distributions differ significantly. This is tested by checking whether enough evidence exist to reject the claim that these distributions are identical. If this was true any of the values contained in either \mathbf{z} or \mathbf{y} are equally likely to be generated from either distribution. The observations in \mathbf{z} and \mathbf{y} can thus be re-allocated, to form new \mathbf{z}^* and \mathbf{y}^* samples that under the null hypothesis should have a similar value for the statistic in question than was obtained for the original samples. Repeating this re-allocation a large number of times a distribution is obtained for the test statistic. If the value of the test statistic evaluated for the original data falls in the tails of this distribution, it is deemed significant and the null distribution is rejected. These steps are summarized by Good (2000) as follows:

- i) Analyse the problem
- ii) Choose a test statistic
- iii) Compute the test statistic for the original samples
- iv) Re-allocate these observations to samples and recompute the test statistic for these rearranged samples
- v) Repeat until the distribution of the test statistic is obtained for all possible permutations
- vi) Accept or reject the hypothesis using this permutation distribution.

Repeating this procedure a large number of times, a subset of the possible permutations of the original values are obtained. The distribution of all the possible re-allocations is known as the permutation distribution and is computed by forming all the possible permutations of the pooled data vector and then allocating first n of them to the one sample and the rest to the other sample. The permutation test does not make use of the entire permutation distribution, but takes a large sample of all possible permutations to approximate the true permutation distribution. The permutation tests above make use of the two sample methodology, which relies on the permutation lemma that states that if a sample of size n is compared to one of size m , the $\frac{N!}{n!m!}$ different permutations (where $N = n + m$) each have an equal probability of occurrence, (Efron & Tibshirani, 1993). The reason for not using the true permutation distribution is that in most scenarios the number of possible permutations becomes prohibitively large. Consequently, the permutation distribution will be approximated when performing the permutation tests.

5.2 APPLICATION

Since the assumption of symmetry is satisfied, permutation tests may be applied. Only the two-sample test procedure is implemented in order to comply with this required assumption. This is applied to the sensory data, tested previously using standard sensory tests. The main interests here are whether the test and control pieces are generated by the same distribution and similarly for the control pieces used in the two separate trials. In the bootstrap tests it was shown that the results for the one-sample and two-sample t -tests did not differ drastically. Not taking the relationship between an individual's test and control scores into account did not seem to have a substantial effect on the obtained results.

The choice of test statistic is the next concern. As the main issue is whether the means of the two distributions differ significantly, a suitable test statistic would thus be the difference in the means. Since the only concern is that the difference between the taste of the test objects and that of the control should not be larger than the random variation found within the control objects, a one sided alternative is appropriate.

The R.function **permutation.test()** was constructed in order to implement the permutation testing procedure. This function consists of the following algorithm (Efron & Tibshirani, 1993, algorithm 15.1):

1. Calculate the test statistic $\hat{\theta}$ for the original two samples \mathbf{z} and \mathbf{y} .
2. Reallocate the observations in \mathbf{z} and \mathbf{y} to obtain \mathbf{z}^* and \mathbf{y}^* ; calculate the permutation replicate of the test statistic, $\hat{\theta}^*$.
3. Repeat 2 a large number of times (in the applications reported here 2000 replicates were evaluated).
4. Compute the approximate ASL_{perm} by calculating the proportion of $\hat{\theta}^*$ -values that is larger or equal to $\hat{\theta}$.

The above procedure is used to test when a one-sided alternative is considered. If a two-sided alternative hypothesis is of concern, the absolute value of $\hat{\theta}^*$ is compared to the absolute value of $\hat{\theta}$. Both the one-sided and two-sided permutation test procedure can be performed by using **permutation.test()**. The appropriate default option is used.

5.2.1 $H_{0I}: \mu_{T3} = \mu_{C3}$ ¹

The null hypothesis considered is whether the difference between the mean of the test and control objects is sufficiently large to imply that a significant difference exists between the test and control items. As mentioned above, the appropriate test statistic is $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C3}$, where \mathbf{x} represents the vector of values and the T3 and C3 refer to the 3-day test and control data, respectively.

Figure 5.1 illustrates the permutation distribution of $\hat{\theta}$ and the original $\hat{\theta}$ value of 1.611 is indicated by the dashed line. An ASL value of 0.0005 was recorded thus strong evidence exists that μ_{T3} is significantly larger than μ_{C3} .

¹ ASL-values and permutation distribution are obtained by applying the R-function **permutation.test()** included in the appendix to the appropriate data sets with **one.sided=T**.

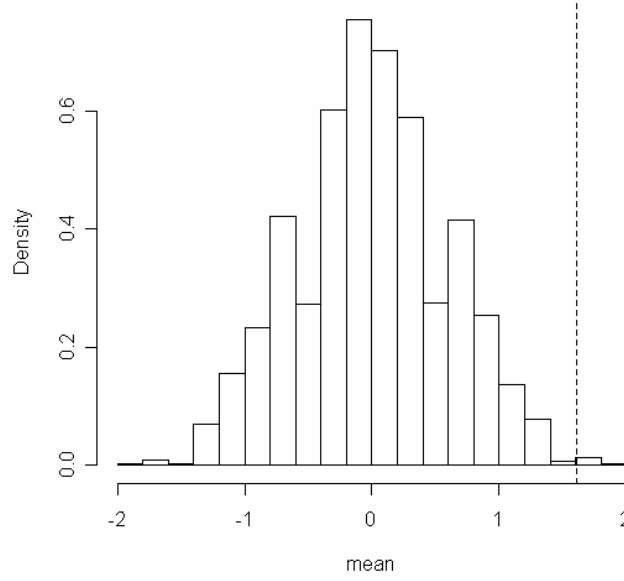


Figure 5.1: Permutation distribution of $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C3}$.

5.2.2 $H_{02}: \mu_{T9} = \mu_{C3}$ ⁹

The next hypothesis considered compares the distribution of the control in the 3-day trial to that of the treatment in the 9-day trial. If it is assumed that the control objects are identically distributed, comparing this result with the one obtained when comparing the treatment and the control objects in the 9-day experiment, will shed some light on the differences between the two panels.

The hypothesis being tested now is whether the mean of the test values for the 9-day data is significantly larger than that of the control data of the 3-day data set. The appropriate test statistic now is $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C3}$. An ASL value of 0.0025 was recorded for $\hat{\theta} = 0.944$ and thus very strong evidence exist that the null hypothesis may be rejected and therefore that the mean of the 9-day test data is significantly larger than that of the 3-day control data. This was once again a one sided hypothesis test and Figure 5.2 illustrates the permutation distribution of the test statistic.

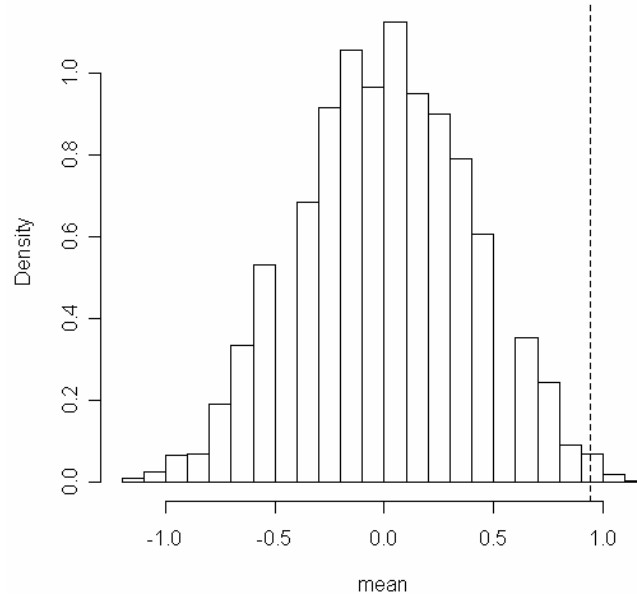
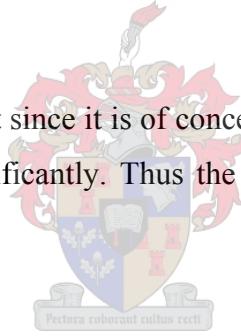


Figure 5.2: Histogram approximating the permutation distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C3}$.

5.2.3 $H_{03}: \mu_{C9} = \mu_{C3}$ ²

This is a two sided hypothesis test since it is of concern whether the mean of the control data of the two trials differ significantly. Thus the alternative hypothesis now is that they are unequal: $H_{a3}: \mu_{C9} \neq \mu_{C3}$.



The test statistic computed to test this hypothesis is $\hat{\theta} = \bar{x}_{C9} - \bar{x}_{C3}$ but since this is a two sided hypothesis test $\hat{\theta} = \bar{x}_{C3} - \bar{x}_{C9}$ would have yielded the same results. The histogram in Figure 5.3 of the test statistic contains gaps, similar to those observed in the bootstrap distribution. Whether these gaps occur due to chance or if the values cannot be obtained due to the limited number of distinct control values allocated, necessitates some further scrutiny. The null hypothesis however cannot be rejected since the ASL value obtained is 0.4715. Therefore it seems that there is no statistical difference between the means of the control data and furthermore that the two panels allocated similar values in the two separate trials.

² Histogram and ASL-values obtained by implementing R-function `permutation.test()` with `one.sided=F`.

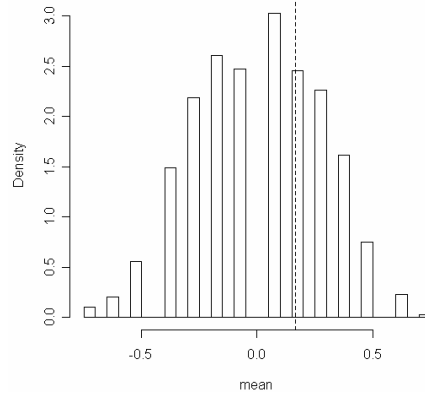


Figure 5.3: Histogram approximating the permutation distribution of $\hat{\theta} = \bar{x}_{C9} - \bar{x}_{C3}$.

5.2.4 $H_{04}: \mu_{T9} = \mu_{C9}$

The hypothesis considered in the 9-day trial is whether the means of the test and control values differ significantly. A one sided alternative is used, that the mean of the test data is significantly larger than that of the control data, i.e.: $H_{a4}: \mu_{T9} > \mu_{C9}$. The test statistic is $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}$ and Figure 5.4 contains the histogram of the approximate permutation distribution of the $\hat{\theta}$ values.

The original value of $\hat{\theta}$ is 0.778 which is significant with an ASL value of 0.0155. Therefore the null hypothesis is rejected and it is concluded that the mean of the test data is significantly larger than that of the control data. This corresponds to the results of the Bootstrap test as well as the one-sample t-test. Once again the histogram contains gaps which will be examined further in Section 5.4.

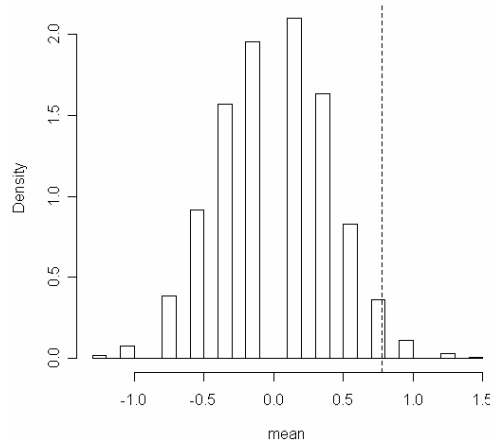


Figure 5.4: Histogram approximating the permutation distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}$.

5.2.5 $H_{05}: \mu_{T3} = \mu_{C9}$

As in Section 5.2.2 the control data from one trial is compared to the test data of another. If the result obtained for the hypothesis that the mean of the test data of the 3-day trial and the mean of the control data of the 9-day trial are similar to that of the permutation test in Section 5.2.1 it also shows that the panels judged the control sets in a similar fashion.

The test statistic considered now is $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C9}$ and the alternative hypothesis is: $H_{a5}: \mu_{T3} > \mu_{C9}$. Figure 5.5 contains the estimated permutation distribution for which an ASL value of 0.0030 is obtained thus the null hypothesis is rejected. It is concluded that the mean of the test data of the 3-day data is significantly larger than that of the control data in the 9-day trial, which corresponds with the result obtained in Section 5.2.1. Therefore there does not seem to be a significant difference between the two panels.

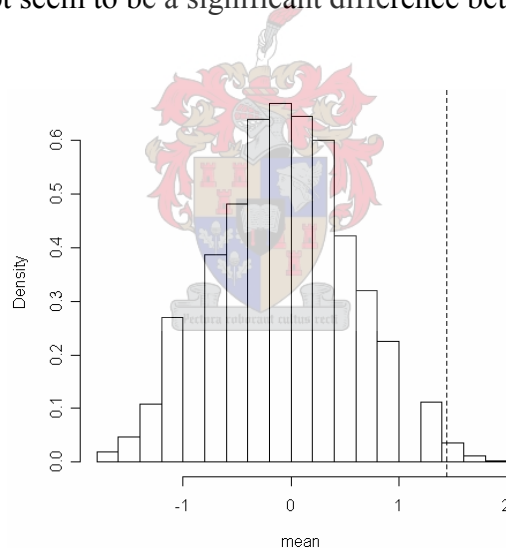


Figure 5.5: Histogram approximating the permutation distribution of $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C9}$ with the dashed line representing the value of $\hat{\theta} = 1.444$.

5.2.6 $H_{06}: \mu_{T9} = \mu_{T3}$

The scores allocated to the test pieces for the two separate trials indicate the perceived difference between the specific test item and control item. To get some idea as to whether the test items in the two trials differ from one another, the two-sided

permutation test is carried out. Keep in mind that some of the panellists participated in both trials, thus the scores may not be totally independent.

The alternative hypothesis for this test is $H_{a6}: \mu_{T9} \neq \mu_{T3}$. The test statistic evaluated is $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{T3}$ and the corresponding histogram of the approximate permutation distribution is displayed in Figure 5.6. The permutation distribution seems symmetrical, but once again contains gaps. The dashed line indicated the observed value of $\hat{\theta} = -0.6667$ with a corresponding ASL of 0.1580 which implies that the null hypothesis cannot be rejected, thus the treatment items in the 9-day trial do not differ significantly from those in the 3-day trial.

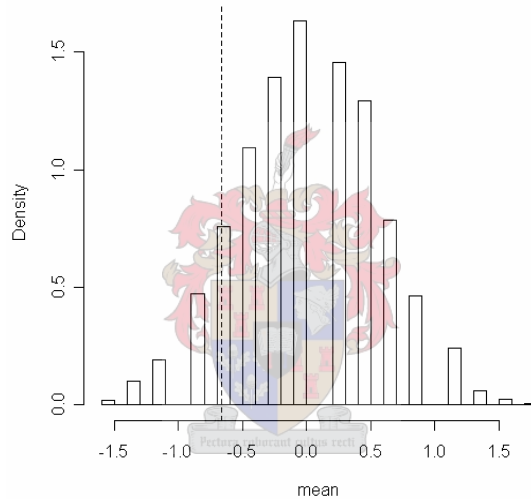


Figure 5.6: Histogram approximating the permutation distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{T3}$ with the dashed line representing the value of $\hat{\theta} = -0.6667$.

5.3 IN CONCLUSION

In the previous section it was shown in all of the tests that for both trials the test values are significantly larger than those of the control. Furthermore, the means of the control data for both trials' data seem to be similar. The scores awarded by the two panels do not appear to differ drastically.

Note however that these tests are implemented in this section as exploratory devices, and not in a strict statistical sense. Adjustments, like implementing Bonferroni's

inequality will have to be implemented and independence of test panels will also need to be examined.

5.4 TRUE PERMUTATION DISTRIBUTION

As mentioned earlier, the permutation test procedure uses the approximate permutation distribution to estimate the true permutation distribution by evaluating a large sample of all the possible permutations. When the histograms were scrutinized, an interesting feature came to light. The histograms contained gaps, thus some values for the mean were never obtained within the sample of 2000 permutations. This raises the question of whether it was possible with the observed values to obtain those values for the mean. The reason that this may be impossible is that when the actual observed values are considered it is noted that several of them are zero and only a few unique non-zero values were recorded.

Since the sample sizes are both only nine it may be possible to calculate the true permutation distribution for each data set. When the 9-day data is considered, there are $\frac{(2n)!}{n!n!} = \frac{18!}{9!9!} = 48620$ possible permutations. Although this is quite a large number, owing to the advances in computer speed, it is possible to compute each of the permutations and determine which values for the mean are possible.

The R-function **real.perm.dist()**³ was constructed to compute the true permutation distribution. This procedure employs the R-package *combinat* to calculate every possible sample of 9 from the 18 observations and then allocates the 9 observations sampled to the first sample and the rest to the second sample. For each of these possible samples, the test statistic is calculated and these replicates form the true permutation distribution.

Figure 5.7 contains the true permutation distribution for $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}$. There were only 14 unique values observed, namely $(-1.444, -1.222, -1.000, -0.778, -0.556, -0.333, -0.111, 0.111, 0.333, 0.556, 0.778, 1.000, 1.222, 1.444)$. Thus the gaps obtained in the

³ The permutation distributions and the histograms were obtained by implementing R-function, **real.perm.dist()** contained in the appendix.

estimated distribution were not just there by chance but is an actual feature of the underlying distribution.

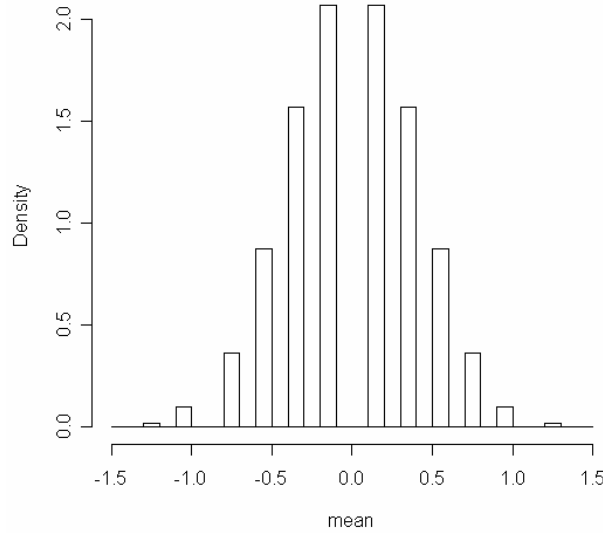


Figure 5.7: Histogram of the true permutation distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}$.

When looking at the permutation distributions of the other test statistics evaluated in the previous section, similar results were also obtained for their true permutation distribution. The permutation distribution of the test statistic for the 3-day data, shown in Figure 5.8, does not contain any gaps, due to the fact that there were 38 distinct values for the test statistic ($\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C3}$), but only 25 intervals used to construct the histogram.

Figures 5.9 to 5.11 contain similar distributions for all the different test statistics and show that if the 9-day data was used in the evaluation of the test statistic, the histograms may contain some gaps. There are more distinct values for the test statistic for each of these distributions, (26, 38, 20) respectively. This illustrates another drawback of a histogram: the dependence on the choice of the number of intervals. These histograms were obtained using 25 intervals, but if this number was decreased, the gaps seen in the histograms would not be visible and the test statistic would seem to be continuous. Similarly, an increase in the number of intervals, would lead to even more gaps and the

test statistic would thus appear to be discrete, having a very fixed range of possible values.

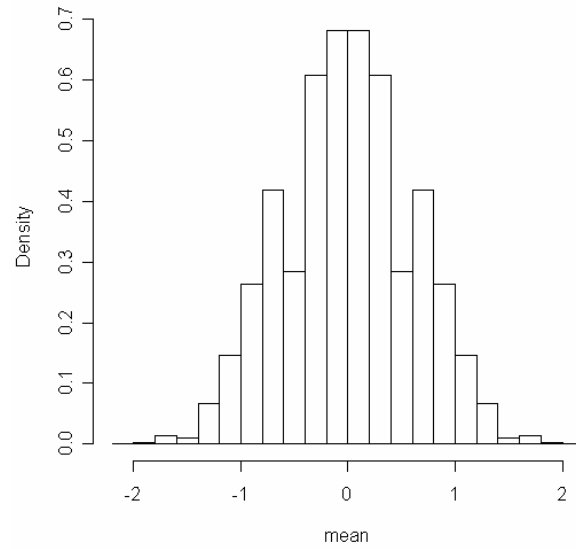


Figure 5.8: Histogram of the true permutation distribution of $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C3}$.

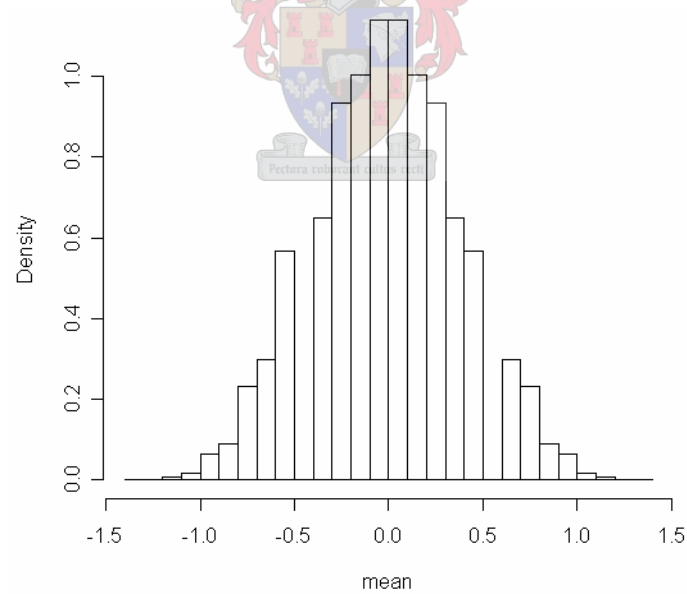


Figure 5.9: Histogram of the true permutation distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C3}$.

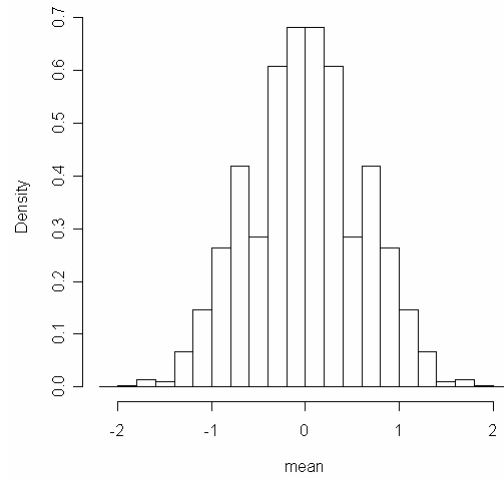


Figure 5.10: Histogram of the true permutation distribution of $\hat{\theta} = \bar{x}_{C9} - \bar{x}_{C3}$.

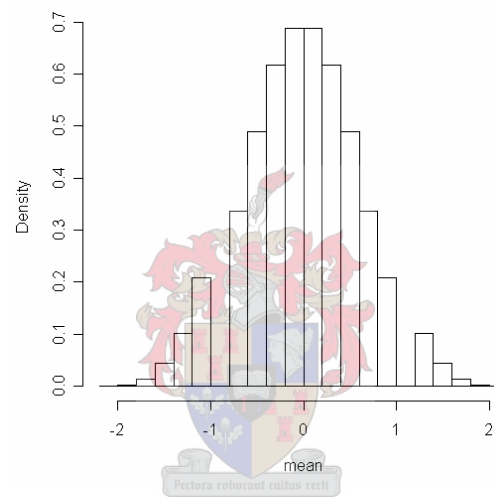


Figure 5.11: Histogram of the true permutation distribution of $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C9}$.

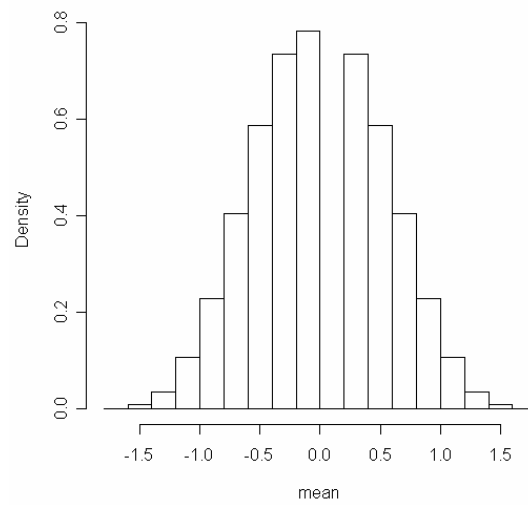
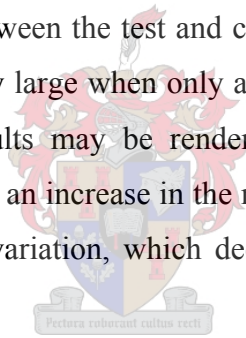


Figure 5.12: Histogram of the true permutation distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{T3}$.

5.5 ACCURACY OF APPROXIMATED PERMUTATION DISTRIBUTIONS

The permutation tests were carried out by taking a sample of size 2000 from the possible 48620 permutations values. This is done since in most cases the total number of permutations is prohibitively large and the calculation of all the possible permutations would be unfeasible. Since the true permutation distribution is available in this example, the stability and reliability of the standard methodology of taking a sample of permutations can now be tested and a minimum number of samples required for an acceptably accurate approximation, may be determined. For every respective number of samples, 1000 iterations of the permutation test are carried out. The number of different theta values as well as the ASL for each of these was recorded.

Table 5.1⁴ consists of the comparison of the true and approximated permutation distribution for the difference between the test and control data for the 3-day trial. The standard error for the ASL is very large when only a small number of permutations are sampled, to the extent that results may be rendered meaningless. The increase in stability of the ASL estimate with an increase in the number of permutations sampled is illustrated by the coefficient of variation, which decreases as the number of samples increases.



In Figure 5.13 the effect of an increase in the number of sampled permutations on the standard error of the obtained ASL's are illustrated. This is done by representing graphically the coefficient of variation as a function of the number of permutations. In Table 5.1 the standard error is estimated using the form given in equation (3.1). It is shown that the ASL only becomes accurate for more than 2000 sampled permutations, due to the elbow in Figure 5.13. Note that for small sample sizes the coefficient of variation is extremely large and thus the estimated ASL-values for such a small number of permutations may be extremely unreliable.

⁴ The R-function `accuracy.perm.test()` contained in the appendix was constructed to compare the true and approximate permutation distributions and compute the values displayed in this table.

Table 5.1.: Comparison between approximated and actual permutation distribution for

$$\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C3}.$$

Number of samples in permutation test	100	200	400	1000	2000	4000	8000
True number of distinct $\hat{\theta}$ values	38	38	38	38	38	38	38
Average number of distinct $\hat{\theta}$ values in permutation tests	23.678	26.485	28.84	31.2890	32.8940	34.1240	34.9870
True Significance Level (TSL)	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011
Average ASL	0.0010	0.0011	0.0011	0.0011	0.0012	0.0011	0.0011
Standard error in ASL estimate	0.0032	0.0023	0.0017	0.0011	0.0007	0.0005	0.0004
Coefficient of variation	3.2000	2.0910	1.5455	1.0000	0.5833	0.4545	0.3636

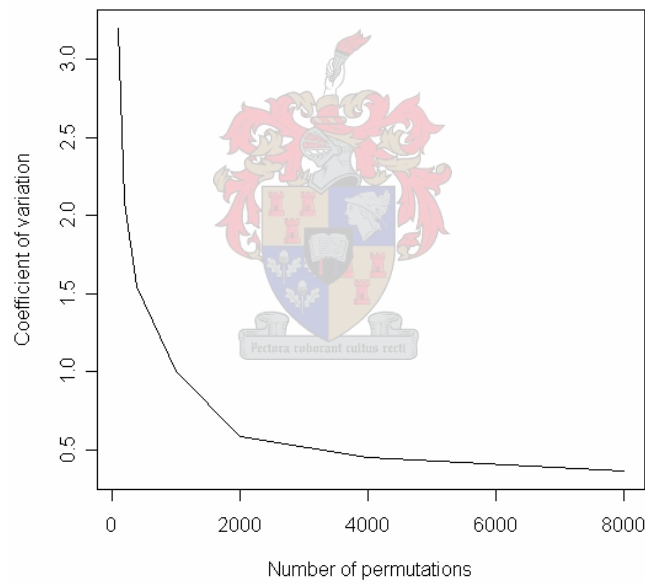


Figure 5.13: Graph to illustrate the relationship between the number of permutations samples and the coefficient of variation for $\hat{\theta} = \bar{x}_{T3} - \bar{x}_{C3}$.

Figure 5.14 represents the true permutation distribution as well as the approximate permutation distribution when 2000 permutations are sampled. This illustrates that the main errors in the approximate permutation distribution lie in the middle of the distribution and not in the tails.

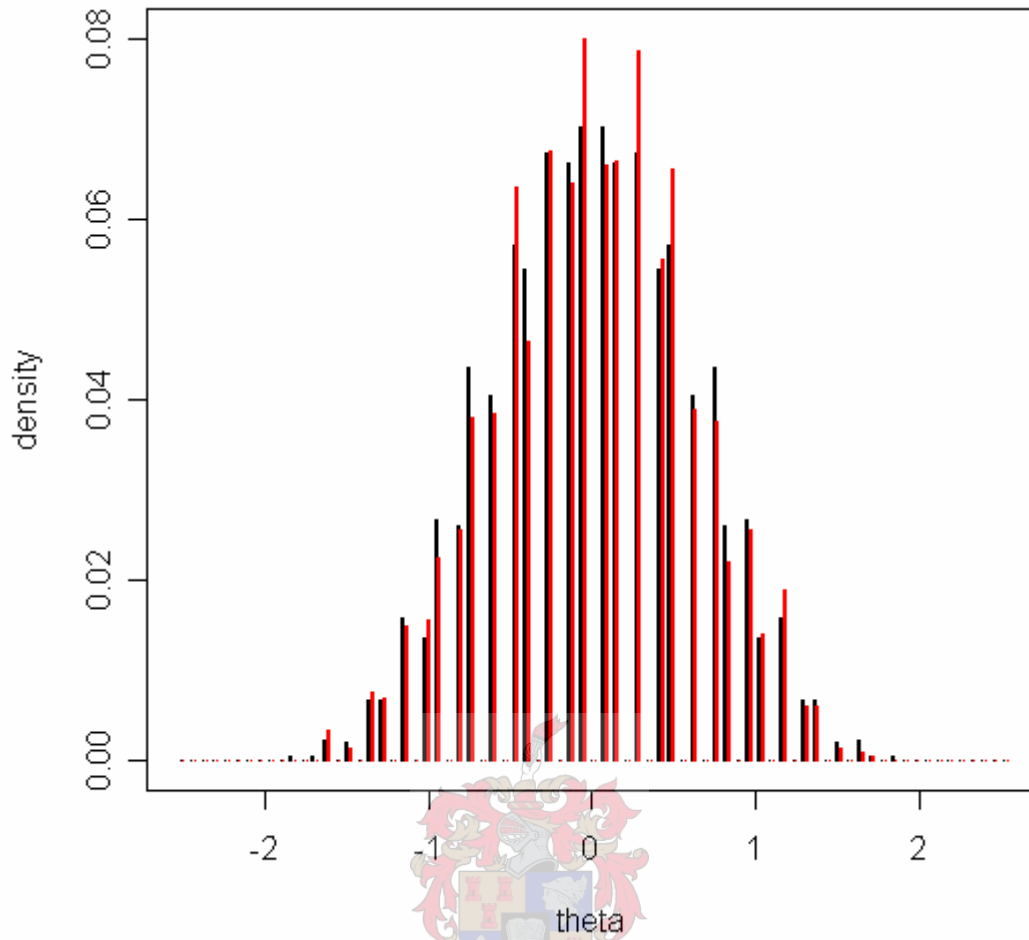


Figure 5.14: Graph comparing the approximate (red) and true permutation (black) distribution of $\hat{\theta} = \bar{x}_{T_3} - \bar{x}_{C_3}$.

Table 5.2¹² contains the values when the approximated permutation is compared to that of the true permutation distribution for the 9-day data when $\hat{\theta} = \bar{x}_{T_9} - \bar{x}_{C_9}$. The average ASL-values are contained in Table 5.2 as well as the standard error of the 1000 ASL-values. The coefficients of variation are much smaller than those contained in Table 5.1. This implies that fewer permutation samples are required to obtain a trustworthy result.

Table 5.2.: Comparison between approximated and actual permutation distribution for

$$\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}.$$

Number of samples in permutation test	100	200	400	1000	2000	4000	8000
True number of distinct $\hat{\theta}$ values	14	14	14	14	14	14	14
Average number of distinct $\hat{\theta}$ values in permutation tests	9.5780	10.6320	11.0500	11.8240	12.3380	12.6560	13.11
True Significance Level (TSL)	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117
Average ASL	0.0114	0.0118	0.0118	0.0113	0.0118	0.0118	0.0117
Standard error in ASL estimate	0.0107	0.0076	0.0052	0.0033	0.0024	0.0018	0.0012
Coefficient of variation	0.9386	0.6441	0.4407	0.292	0.2034	0.1525	0.1026

It is apparent that for 2000 or less sampled permutations, all the distinct values for theta will not be attained which might lead to misleading results. As the number of samples increases, this problem is eliminated. The ASL-values are however the main concern since they are used in hypothesis testing. These values also seem to become more stable as the number of samples increases. The variation within the ASL-values would only have had an effect at a significance level of 1% or lower. The standard error indicates a cause for concern when less than 1000 permutations are sampled, since the ASL value will still be very unstable due to the fact that the standard error is large relative to the mean.

Figure 5.15 illustrates how the stability increases with the number of permutations. Again it is clear that the ASL-values only become reliable when the number of permutation samples exceeds 1000. It is important to keep the scale of the graph in mind, since although the shape of the figure is very similar to that of Figure 5.13, the range of the coefficient of variation is much smaller.

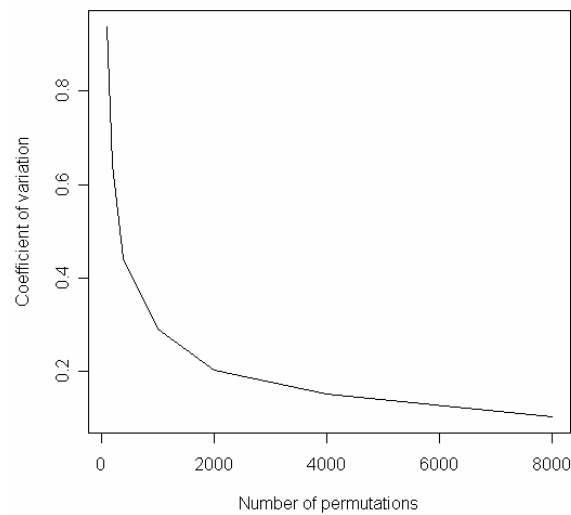


Figure 5.15: Graph to illustrate the relationship between the number of permutations samples and the coefficient of variation for $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}$.

Figure 5.16 illustrates the difference in the histograms of the approximated and true permutation distribution when 2000 permutations are sampled. It is interesting to note that most of the discrepancies lie in the middle of the distribution, and not in the tail, which is quite reassuring when hypothesis testing is concerned.

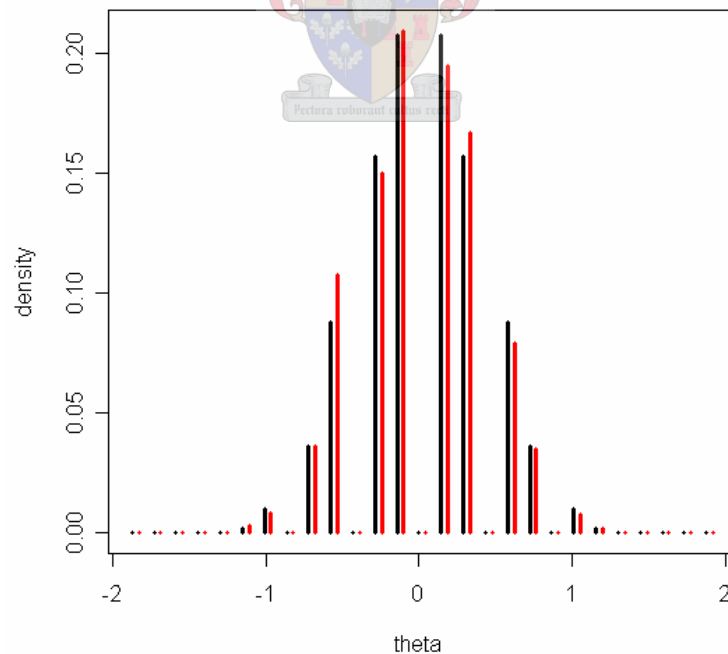


Figure 5.16: Graph comparing the approximate (red) and true permutation (black) distribution of $\hat{\theta} = \bar{x}_{T9} - \bar{x}_{C9}$.

Table 5.3 contains the result when evaluating the accuracy of the permutation test when testing whether a significant difference exists between the control items of the two trials. Since the test was not significant, the error in the ASL does not influence the obtained results, but the error actually is quite large and may have played a role had the TSL been significant. All the different possible theta values are obtained for very few sampled permutations and the average ASL value converges quite rapidly.

Table 5.3.: Comparison between approximated and actual permutation distribution for

$$\hat{\theta} = \bar{x}_{C9} - \bar{x}_{C3}.$$

Number of samples in permutation test	100	200	400	1000	2000	4000	8000
True number of distinct $\hat{\theta}$ values	14	14	14	14	14	14	14
Average number of distinct $\hat{\theta}$ values in permutation tests	12.005	12.997	13.638	13.971	14	14	14
True Significance Level (TSL)	0.2285	0.2285	0.2285	0.2285	0.2285	0.2285	0.2285
Average ASL	0.2290	0.2281	0.2280	0.2280	0.2285	0.2284	0.2287
Standard error in ASL estimate	0.0422	0.0297	0.0210	0.0128	0.0092	0.0066	0.0047
Coefficient of variation	0.1843	0.1302	0.0921	0.0561	0.0403	0.0289	0.0206

The average ASL in Figure 5.17 seems to converge much faster than those in the previous cases, since the coefficient of variation declines much faster. Once again it is shown that the variability of the ASL-values decreases as the number of permutations sampled increases and thus a large number of permutations might be needed. The shape of the line is similar to those displayed in Figures 5.13 and 5.17.

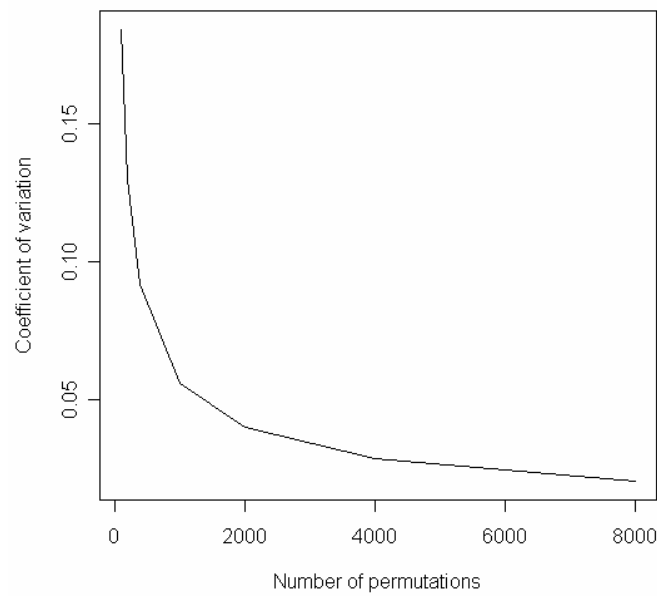


Figure 5.17: Plot to illustrate the convergence of the ASL-values when $\hat{\theta} = \bar{x}_{C9} - \bar{x}_{C3}$.

Figure 5.18 illustrates the fact that the approximation is reasonably accurate for this test statistic. Except for one theta value, there does not seem to be any noteworthy differences between the approximate and true permutation distributions.

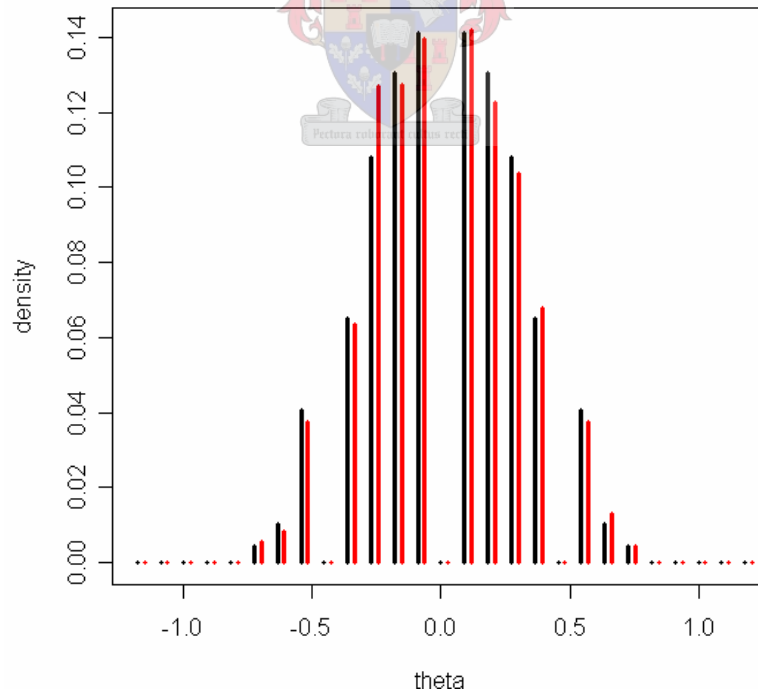


Figure 5.18: Graph comparing the approximate (red) and true permutation (black) distribution of $\hat{\theta} = \bar{x}_{C9} - \bar{x}_{C3}$.

In Figures 5.13, 5.15 and 5.17 it is seen that the plot of the coefficient of variation forms an elbow at approximately 2000 permutations, thus it is recommended that at least 2000 permutations need to be sampled to obtain consistent results. Furthermore, the mean of the ASL-values in each of the tables in this section seem to be quite stable. This can be attributed to the fact that performing 1000 iterations of a permutation test, each sampling 200 (for example) of the possible permutations, or just sampling 200000 replications will deliver the same means ASL if the sampling is random.

5.6 ASSOCIATION IN PANEL SCORES

The performance of panellists is crucial in sensory science since this determines the reliability of the results. Expert panels were used in the taint tests and one would expect their judgement to be consistent. Thus, due to their extensive training, it is assumed that they will award the same score when confronted with identical objects at different points in time. The scores allocated to the control item by a specific panellist in each trial should therefore be identical or at least very similar. Establishing whether a correlation between these two sets of scores exists, will now be considered. Only the scores from the panellists that participated in both trials will be used.

Since it has been found that the parametric assumptions are not met by the 3-day and 9-day data sets, non-parametric correlation measures will be used to quantify the association of the allocated scores. Note that determining whether scores are related differs from the process of determining to what extent they are associated. The probability of observing as extreme a value for the correlation, under the null hypothesis that scores are unrelated, can also be estimated.

The sample correlation coefficient (r) can be evaluated to describe this association and Pearson's product-moment correlation coefficient is often used (Siegel, 1956). Statistical inference regarding r however depends on normality assumptions and thus is not appropriate for the study at hand. The contingency coefficient is a non-parametric measure of correlation that does not rely on any distributional assumptions and is relatively simple to implement. The only necessary assumption is that the underlying scores, which can be viewed as categorical, are continuous. A contingency table is

created from the two sets of scores and the observed frequencies in each cell calculated. These frequencies are compared to the expected frequencies computed under the assumption that there is no association between the scores. If the discrepancies between the observed and expected frequencies are large, the value of r will be high.

The contingency table for the control scores under consideration is displayed in Table 5.5. Since there is only a single category for the scores in the 3-day trial a true contingency table cannot be computed and thus the contingency coefficient cannot be applied to this problem.

Table 5.5: Contingency table for the control data for the to trials

		9-day trial	
		0	1
3-day trial	0	4	2

The Spearman rank correlation coefficient (r_s) considers only the rank of the scores and not the actual values. Siegel (1956) contains all the details on how to calculate the Spearman rank correlation coefficient as well as illustrative examples.

Table 5.6 contains the appropriate ranks for the data under consideration. When the scores are tied, the mean value of all the ranks that would have been allocated to the tied items, is allocated to each of them. Thus for the 3-day scores the appropriate rank is $\sum_{j=1}^6 j / 6 = 2.6667$. When a large proportion of ties are recorded a correction factor is incorporated in the computation of r_s . As can be seen in Table 5.6 the data under consideration contains a large number of ties, thus the correction factor used to compute r_s becomes so large that r_s is no longer reliable.

Table 5.6: Appropriate quantities for the 3-day and 9-day control scores.

	3-day scores	x_i	9-day scores	y_i	d_i	d_i^2
i	0	2.6667	1	5.5	-2.8333	8.0278
iii	0	2.6667	1	5.5	-2.8333	8.0278
iv	0	2.6667	0	3.25	-0.5833	0.3403
vii	0	2.6667	0	3.25	-0.5833	0.3403
viii	0	2.6667	0	3.25	-0.5833	0.3403
ix	0	2.6667	0	3.25	-0.5833	0.3403
Total					-8	17.4167

When using the calculations summarised in Table 5.6 to compute r_s , it is found that the large number of ties results in a denominator of zero, and r_s can therefore not be calculated. The Kendall rank correlation coefficient was also considered, but since only two scores were allocated by the panellists, this statistic would not be appropriate.

Figure 5.20 illustrates the change in score allocated to the control items in the two separate trials. It can be seen that most of the judges were consistent and accurate, awarding zeros to both the control scores. Judges (i) and (iii) did however award a slightly larger value to the control items in the 9-day trial. If their scores differed radically from the scores awarded in the 3-day trial, it might indicate that some further training is necessary for these two expert panellists. An important fact to keep in mind is that the MQM test procedure deems a difference of one to be significant and thus the scores of one awarded by judges (i) and (iii) would be very influential.

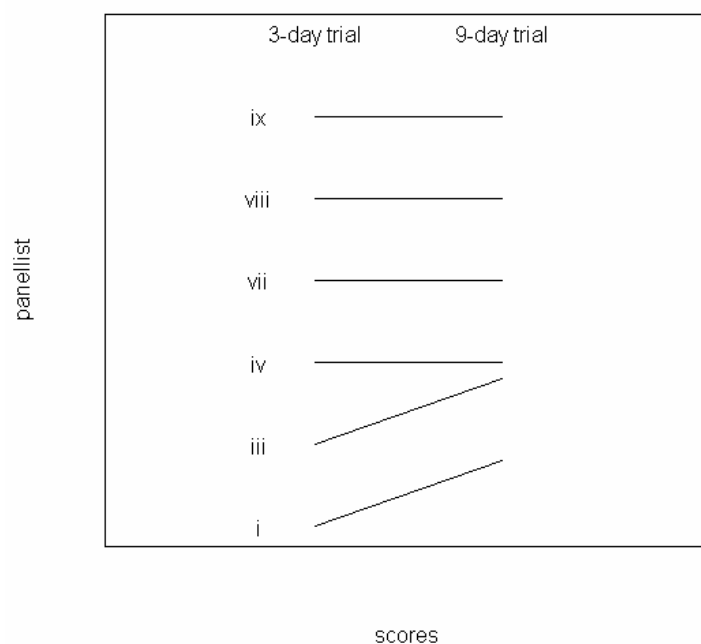


Figure 5.19: Schematic representation of the allocated scores.

To conclude...

When comparing the results obtained by the t-test for testing whether a significant difference exists between the test and control data to that of the permutation tests for either trial the results are similar. Thus implementing either of these procedures would lead to the same conclusion in both trials, namely that a significant difference exists between the test and control items.

Furthermore, when testing whether the control items used in the two trials differed, an insignificant result was obtained. This is satisfactory, since the same items were used in both trials, but it illustrates an important fact: the two panels used in the separate trials awarded similar scores to the control items, and thus panels' judgements seem to be consistent. This consistency is observed for a specific unchanged item but how consistent are panellists when it comes to experiencing differences in products? This leads to the question of how sensitive the panel is to changes in the composition of products. In the following chapter results obtained in sensory discrimination trials will be compared to the biplot representation of the composition of a product before and after modifications to the standard product have been made.

CHAPTER 6

COMPARING RESULTS FROM SENSORY TESTS WITH BILOTS

Sensory discrimination trials are now studied. For this, the question under consideration is whether large distances between the composition of objects obtained by a multidimensional technique such as biplots would correspond to significant results obtained in sensory trials. If this is true, identifying which ingredient in a product plays the decisive role that would cause a panellist to be able to discriminate between two products, may be possible. The use of sensory trials for a specific product may be rendered superfluous if the ingredient and the amount of change necessary to discriminate between items can be identified in this way.

6.1 AN INTRODUCTION TO BILOTS

Biplots are graphical displays that aim to represent multidimensional data in a low dimensional space and to describe both the variables and samples in one display (Gower, 1996). Difficulties encountered when trying to visualise multidimensional data are addressed by approximating the distances between objects by some form of multidimensional scaling. This approximation may then be projected onto a two or three dimensional space that can be graphically displayed so that the differences between objects may be visually inspected. The variables that generated this display are now represented by axes in a similar fashion to traditional Cartesian axes and can be used to attain the value of a specific variable for a specific object by orthogonal projection onto the relevant axis. This set of axes is called the prediction axes. The set of axes used to place a new object on to the original display is known as the interpolation axes, and are generally not the same as those used for prediction.

Multidimensional data often originate from inspecting n objects or samples with regards to p variables or attributes (Cox & Cox, 2001). These measurements may be nominal, ordinal or continuous numerical values. An important consideration is the choice of a distance measure used to quantify the differences between objects. This metric depends on the type of variables under consideration and may be Pythagorean, Mahalanobis or Chi-Square distances amongst others.

6.2 Principal component analysis (PCA) biplots

PCA is a technique which reduces the dimensionality of a data set by considering the largest eigenvalues associated with the matrix $\mathbf{X}'\mathbf{X}$, where \mathbf{X} is the data matrix. The goal of our representation is to portray the intersample distances between the samples accurately in some sense. In multidimensional scaling a measure of the inaccuracy is defined and the objective now is to minimise this error (Cox & Cox, 2001).

These n multidimensional measurements may be represented in p -dimensional space \mathcal{R}^p by utilizing p Cartesian axes. If the data are gathered in a matrix $\mathbf{X}: n \times p$, the rows of this matrix form the coordinates of the n samples in p -dimensional space. The subspace \mathcal{L} of \mathcal{R}^p is the r -space chosen by the method of principal components that minimises the least squares error criterion viz. $tr\left\{(\mathbf{X} - \mathbf{X}^*)(\mathbf{X} - \mathbf{X}^*)'\right\}$, where \mathbf{X}^* is the least squares estimator for \mathbf{X} (Cox & Cox, 2001). This space is spanned by the eigenvectors of $\mathbf{X}'\mathbf{X}$ associated with the largest r eigenvalues. Let \mathbf{V}_r be a matrix consisting of these r eigenvectors as columns. The coordinates of the n samples in the subspace \mathcal{L} spanned by the columns of \mathbf{V}_r are given by $\mathbf{Z} = \mathbf{X}\mathbf{V}_r$. To find the coordinates of a new sample with measurements \mathbf{x} , these measurements need to be orthogonally projected onto \mathcal{L} which is obtained as $\mathbf{z}' = \mathbf{x}'\mathbf{V}_r$. Axes are obtained by noting that \mathbf{V}_r represents the rotation of the original Cartesian axes to the principal axes (cf. Gower & Hand, 1996).

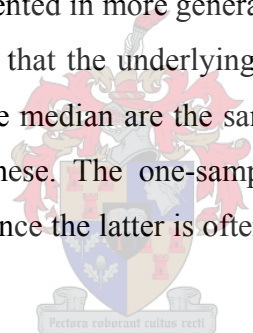
Before applying this methodology to the composition of the products under scrutiny, the procedures used to test whether a significant difference between the attributes of these products exists, will receive some attention.

6.3 PROCEDURE FOR TESTING PAIRED COMPARISON DATA

6.3.1 Introduction to the Wilcoxon paired sample test

The Wilcoxon paired sample test is a nonparametric analogue to the paired-sample t-test. This test procedure is also referred to in the literature as “matched pairs” with “rank sum” or “signed rank” in some combination with “Wilcoxon”.

The paired procedure may be implemented as an alternative to the binomial test and is more powerful (Zar, 1996). Since the Wilcoxon paired test does not make the normality assumption, it can also be implemented in more general circumstances than the t-test. The more relaxed assumption is made that the underlying distribution should be symmetric. This implies that the mean and the median are the same and therefore this methodology may be applied using both of these. The one-sample t-test is more powerful if the normality assumption holds, but since the latter is often not the case, the Wilcoxon paired sample test is often preferred.



The procedure involves calculating the differences in the pairs and assigning ranks from low to high (i.e. assigning rank 1 to the smallest absolute difference and rank n to the largest) to the absolute differences, appending each rank with the sign of the original difference. Zero differences are ignored when assigning ranks. All the positive ranks are summed (T_+ , say) as well as all the negative ranks (T_- , say). For a two-sided test, these totals are compared to the tabulated critical value $T_{\alpha(2),n}$, see for example Siegel (1956), and if either of these is less than or equal to this critical value, the null hypothesis is rejected. The one-sided testing procedure is slightly different, but since only the two-sided case will be used in this application, further details are omitted.

6.3.2 A normal approximation to the Wilcoxon paired sample test

This procedure for approximating the Wilcoxon paired sample test was developed specifically for large data sets, i.e. when a large number of paired comparisons are made (100 or more). In these cases the distribution of T (whether T_- or T_+ is used) may be approximated by the normal distribution, see for example Siegel (1956), with parameters:

$$\mu_T = \frac{n(n+1)}{4}$$

a

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

These are used to calculate the appropriate Z-statistic:

$$Z = \frac{|T - \mu_T|}{\sigma_T}.$$

The latter statistic is used to calculate the corresponding p-value or may be compared to the critical value of the standard normal distribution $Z_{\alpha(2)}$ in the two-sided test and the null hypothesis rejected when the calculated Z exceeds this critical value. A continuity correction should also be employed as follows:

6.4 FOUR STANDARD SENSORY TRIALS

Triangle and paired discrimination tests were performed to ascertain whether modifications in the composition of a product would be detectable. The reasons for these modifications were to lower the cost and complexity of production, to comply with new legislation as well as due to shortages of the necessary raw material. The manufacturers do not want these changes to be noticeable. Expert panels were used for all of the trials and a fixed set of attributes was prescribed with which the objects were judged in the paired discrimination tests.

6.4.1 Results obtained in paired discrimination tests

The objective of these tests was to ascertain whether there were noticeable changes in specific features of the two separate products. In each test each panellist evaluated a fixed set of attributes that was under consideration. They were presented with the standard product as a control object and the modified product as the test item. Numerical scores ranging from zero to five were awarded to each attribute for the control and test item, respectively, where zero correspond to a low value and five to a high value. The differences between the test and control values were calculated and the signed ranks allocated. These were used as described above and the significance levels obtained for the two trials are collected in Table 6.1 by implementing the appropriate normal approximation to the Wilcoxon paired comparison test. The application of the normal approximation to the Wilcoxon paired comparison test is carried out with some apprehension since this approximation was developed for large samples, but the panels under scrutiny here (that serve as the samples in this application) are only of sizes 26 and 21, respectively. The significance levels printed in red correspond to significant differences between the test and the control item.

Table 6.1.: Significance level attained using the Normal approximation to the Wilcoxon test

	Test A	Test B
Attributes	Significance level	
DT	0.0019	0.3066
IM	0.0503	0.7794
F	0.0147	0.5049
DE	0.2735	0.0090
M	0.0012	0.6165
IR	0.1422	0.7897

From these results it is apparent that there was a significant difference in most characteristics between the modified and original product under scrutiny in test A, but this was not the case in that of Test B. PCA biplots may also be applied to this problem to

ascertain whether the distance between the points corresponding to the original and modified product is consistent with the result of the sensory test.

6.4.2 Results obtained in triangle tests

These two trials were also performed at the product development centre mentioned earlier, to determine whether the changes in the composition of two distinct products would be detectable. This is an excellent example of how standard sensory procedures are modified and then implemented in practice. In both these trials there were several sessions where a number of panellists was presented with a sequence of three items, two test items and one control or vice versa, in a specific order. This order was changed in each session. Panellists were asked to identify the item that differed from the other two. The data for these sessions were pooled for each product, such that Trial C consisted of 20 observations and Trial D of 23. The observations were then analysed assuming that the process of identifying the test amongst two controls would be the same as identifying the control item amongst two test items.

Furthermore the fact that some panellists participated in more than one session. This could introduce some dependency within the obtained results which was ignored when the analysis was carried out. Thus assuming these modifications do not have any effect on the independence of the observations and that the null hypothesis that there is no difference between the test and control items is valid, the number of successes is distributed according to $Bi(n, 1/3)$, with $n = 20$ and $n = 23$ for the two tests considered, respectively. Table 6.2 contains the number of successes as well as the corresponding significance levels obtained in these two separate trials. Neither of these trials detected a significant change in the product owing to the change in product composition.

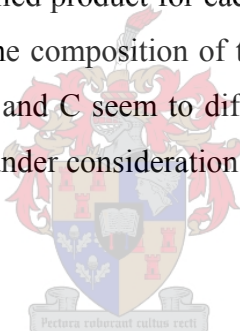
Table 6.2.:Results attained in Triangle tests.

	Test C	Test D
Number of panellists	20	23
Number successes	4	6
Significance level	0.9396	0.8305

6.5 Applying biplots⁸

The four sensory trials were aimed to test whether a change in the composition of four distinct products would be detectable. The two traditional sensory discrimination procedures, namely paired comparisons and triangle tests, were used to address this matter, but were modified when applied. The actual physical compositions of the original as well as the modified products are also available. These will now be used to construct a PCA biplot of the composition of the four original and future products. Whether a large distance between an original and modified product corresponds to a significant hypothesis test is of interest. The significance of the specific ingredients may also be considered.

In Figure 6.1 it is illustrated that the composition of the products in Test B and Test D seem to be very similar. The modified product for each of these trials also does not seem to vary much from the original. The composition of the modified as well as the original products under scrutiny in Test A and C seem to differ drastically. These products also vary a lot from the other products under consideration.



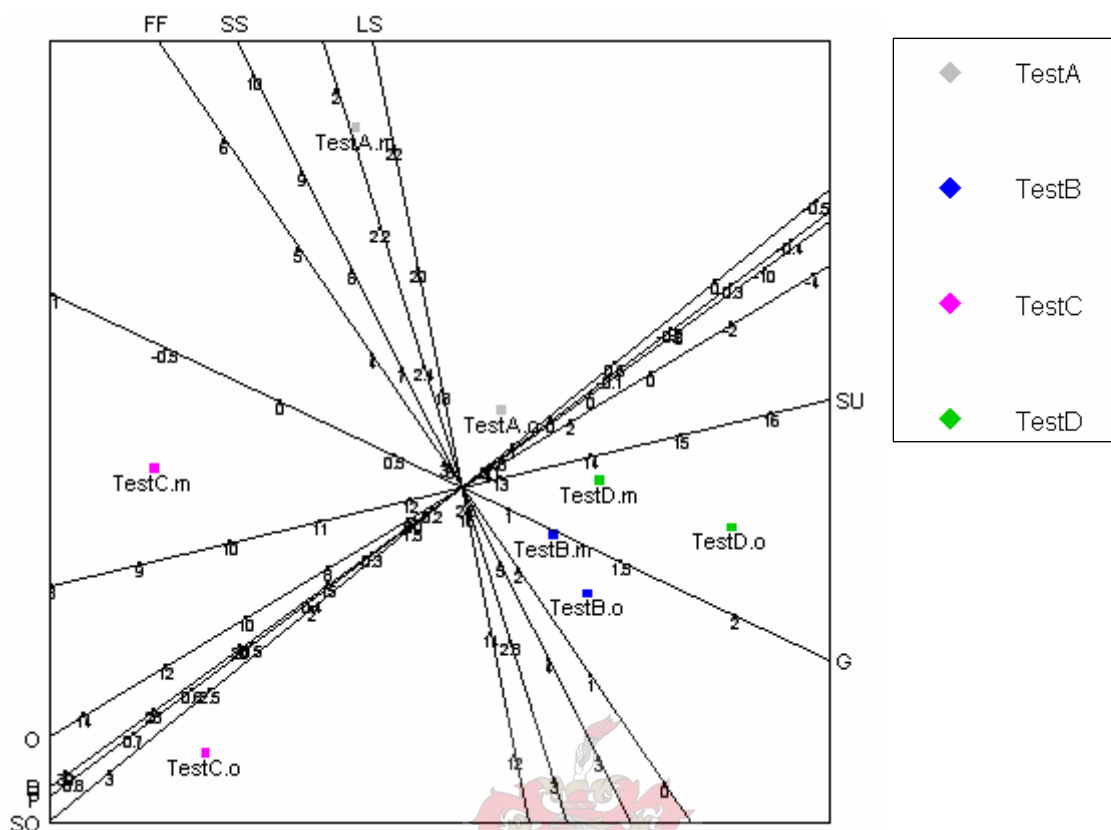


Figure 6.1: PCA Biplot of the composition of products under scrutiny in Test A to D. An “o” postscript refers to the original product and an “m” to the modified version.

In Figure 6.2 colour coded arrows were added to emphasise the shift caused by the change in composition of each product. Table 6.3 contains a colour coded summary of the test results corresponding to the change in composition. In Test A significant differences were observed for most of the attributes. Corresponding to this, a large shift in the position of the product under consideration was illustrated in the biplot, therefore the results for the biplot and significant paired comparison test seem to be consistent in this case. Thus it is concluded that changes in the FF, SS, LS and N ingredients cause significant changes in all but the DE and IR attributes. Furthermore a small shift in these ingredients in Test B, which is represented by the blue arrow, corresponds to an insignificant test in all but one (DE) attribute.

Although Tests C and D both rendered insignificant results, the product shifts illustrated in the biplot differ. The shift in the composition of the product in Test C is large, but that of Test D is much smaller even though the significance level in this test was somewhat lower. As mentioned in Chapter 2, it is known that the power of the paired comparison is better than that of the triangle test (Bower, 1996). This explains why the triangle test does not pick up the shift in the composition of the product in Test C although it is illustrated in the biplot, but the paired comparison test does detect a shift of similar size in Test A. The shift in D is in a different direction than those of the other products, which may have a different effect on the taste of the product.

Table 6.3: Summary of the results of Test A to D

Paired comparison tests		Significance level	
	Attributes	Test A	Test B
	DT	0.0019	0.3066
	IM	0.0503	0.7794
	F	0.0147	0.5049
	DE	0.2735	0.0090
	M	0.0012	0.6165
	IR	0.1422	0.7897
		Test C	Test D
Triangle tests	Number of panellists	20	23
	Number successes	4	6
	Significance level	0.9396	0.8305

It is important to keep in mind however that the biplot considers the data as multivariate and thus considers all the attributes at once, while the Wilcoxon tested for significant differences for each attribute separately. Thus the significance levels computed in the Wilcoxon test do not reflect the significance in difference of the entire product, i.e. all the attributes combined .

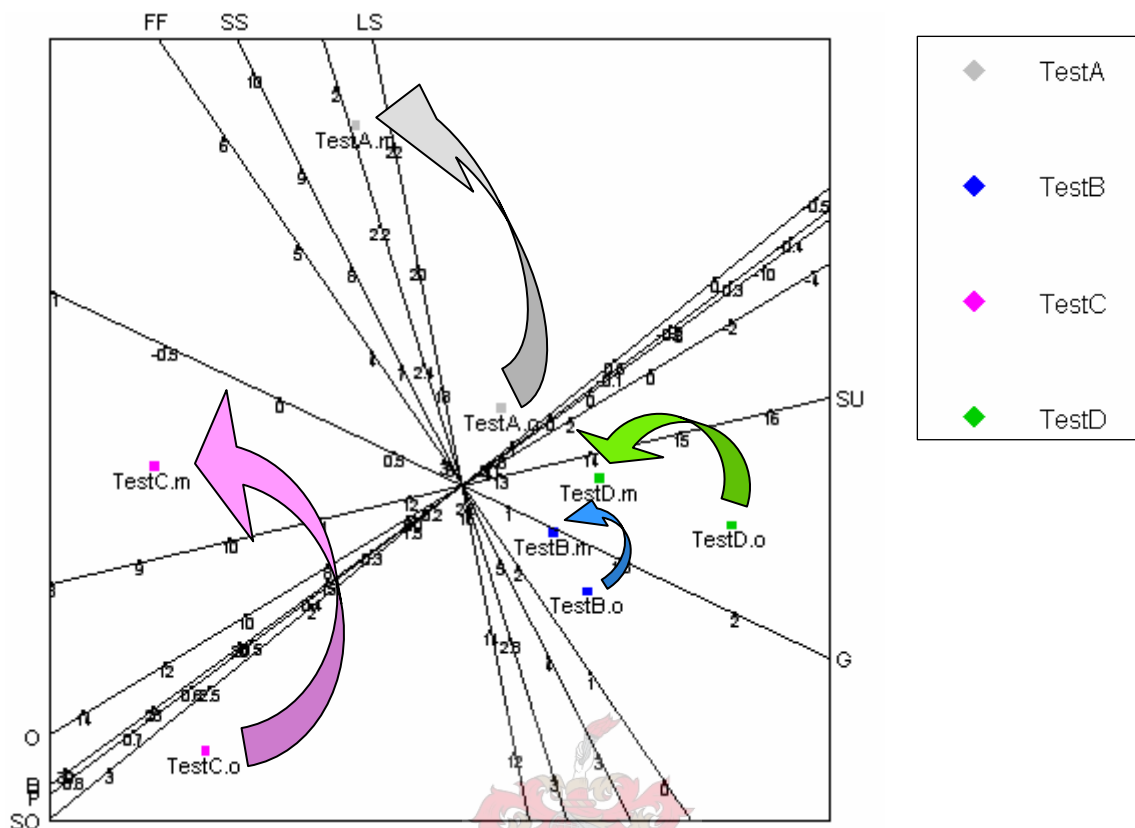


Figure 6.2: PCA Biplot of the composition of products under scrutiny in Test A to D with colour coded arrows added. An “o” post script refers to the original product and an “m” to the modified version.

6.6 INFLUENCE OF PANEL SIZE IN TRIANGLE TESTS¹³

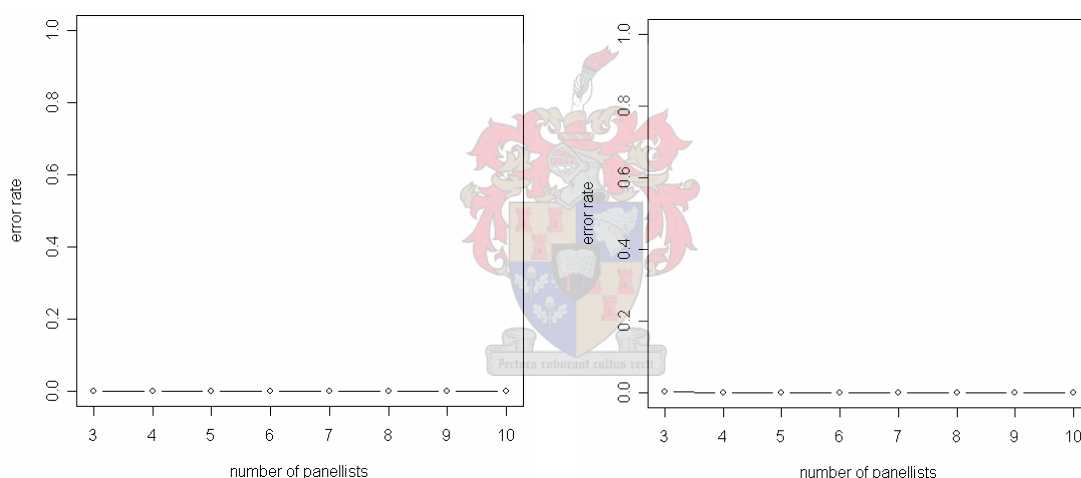
The standard triangle test procedure prescribed by the organisation implies that 10 panellists are sufficient to obtain reliable results. In practice it is often difficult to get more than 5 to 7 panellists for a session. This is an extremely small number and trials carried out in Tests C and D were performed with 20 and 23 panellists, respectively. The influence of a smaller number of panellists will now enjoy some attention.

Since the results simply consist of the number of successes it is simple to “generate” new data. A standard panel size of 20 will be considered and the number of successes will

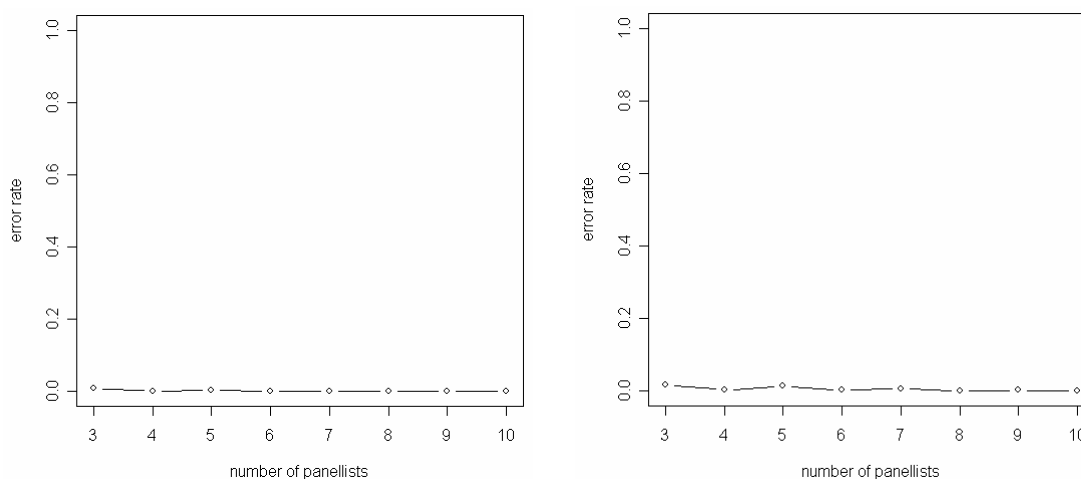
¹³ The figures in this section as well as the error rates contained in Table 6.3 were computed by implementing the constructed R-function `repeat.triangle.test()`.

vary from 1 to 19. For each number of successes, all possible combinations of 3 to 10 panellists, respectively, will be drawn from the complete panel and the result obtained by this reduced panel compared to that of the complete panel. Once again, this does not mean that the result of the complete panel is correct, but this is the most reliable result available against which to judge the smaller panels.

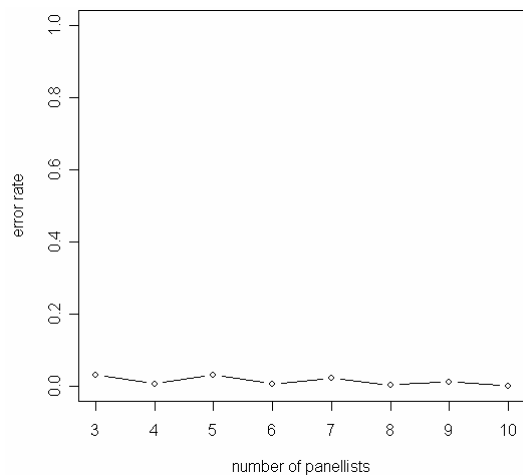
Figure 6.3 a) to r) illustrates the behaviour of the error rate for each respective number of successes. As in Section 3.5 the error rate for a specific panel size is calculated by computing the proportion of the smaller panel's results that differ from those obtained by the complete panel. The zigzag patterns in Figures i) to n) emphasize the role a single panellist can play when the underlying significance level is borderline. These show how crucial the size of a panel can be and illustrate the effect it has on obtained results.



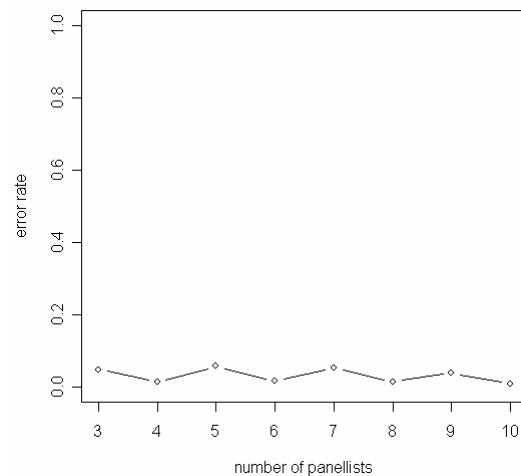
a) Error rate when 3 successes were observed. b) Error rate when 4 successes were observed.



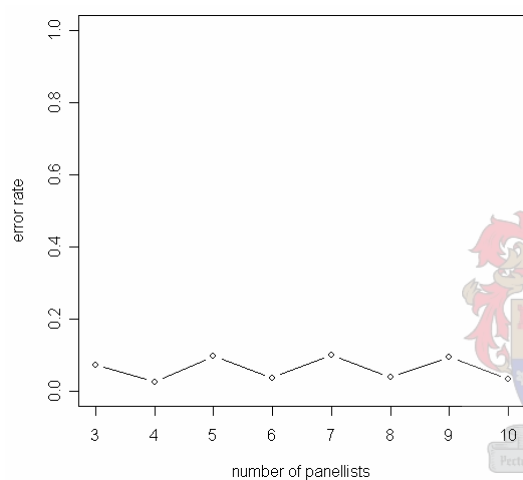
c) Error rate when 5 successes were observed. d) Error rate when 6 successes were observed.



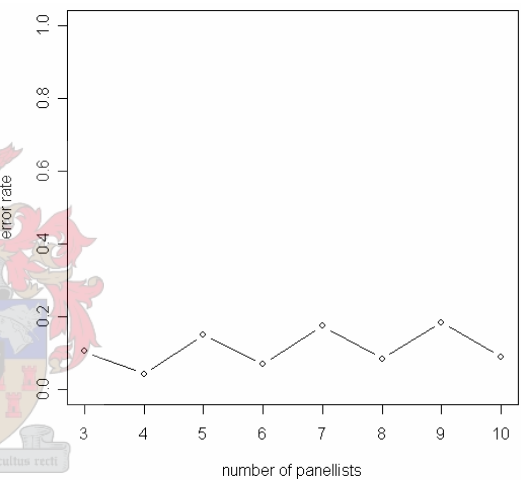
e) Error rate when 7 successes were observed.



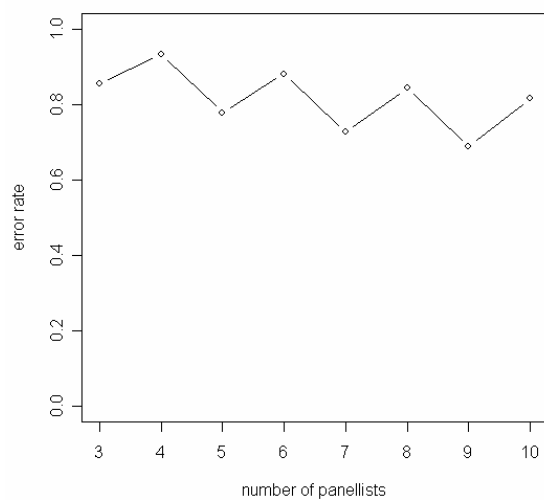
f) Error rate when 8 successes were observed.



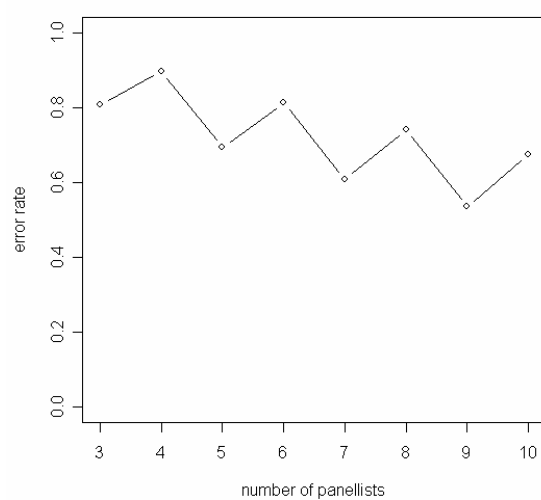
g) Error rate when 9 successes were observed.



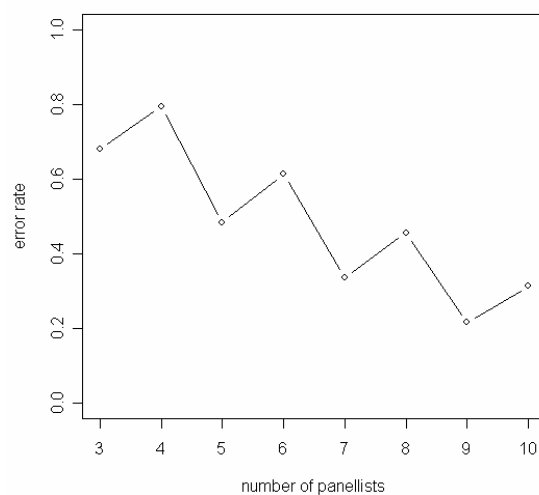
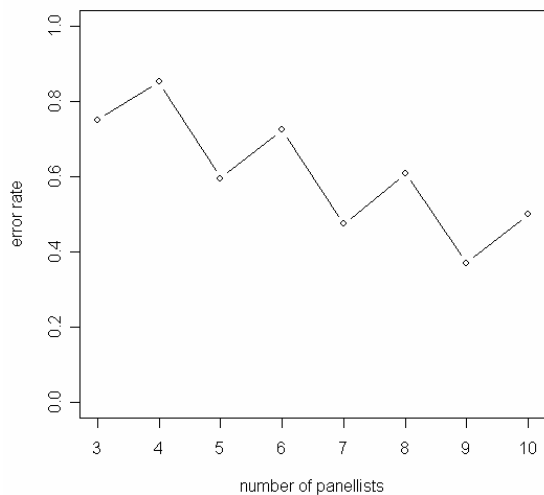
h) Error rate when 10 successes were observed.



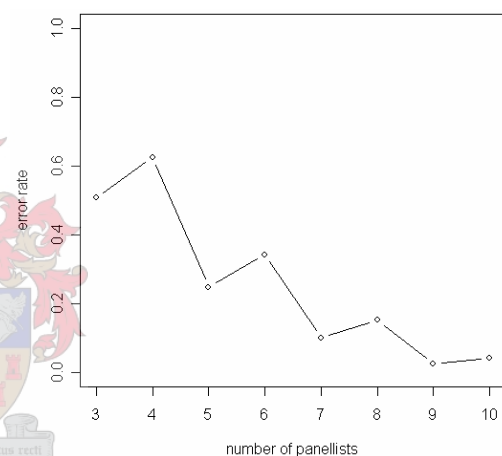
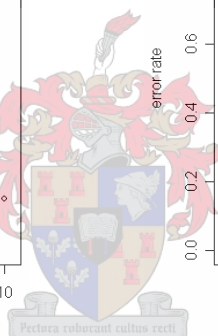
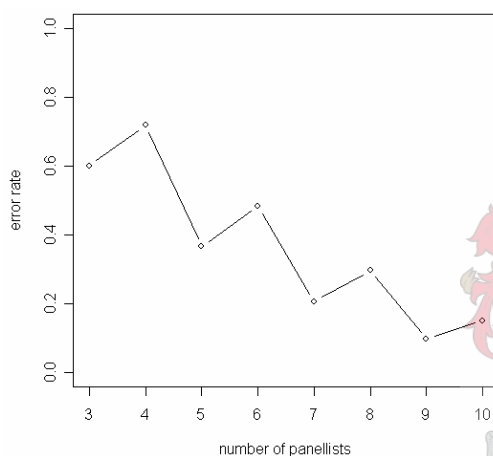
i) Error rate when 11 successes were observed.



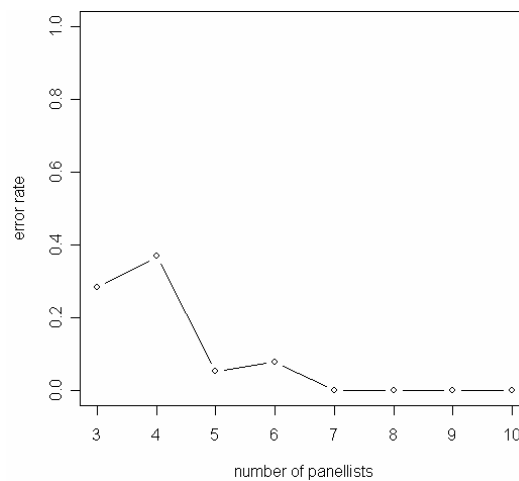
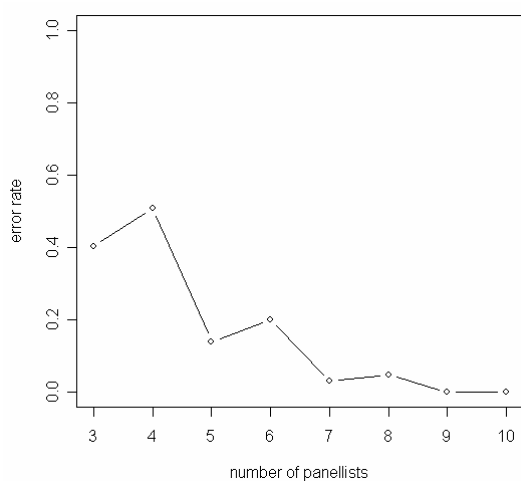
j) Error rate when 12 successes were observed.



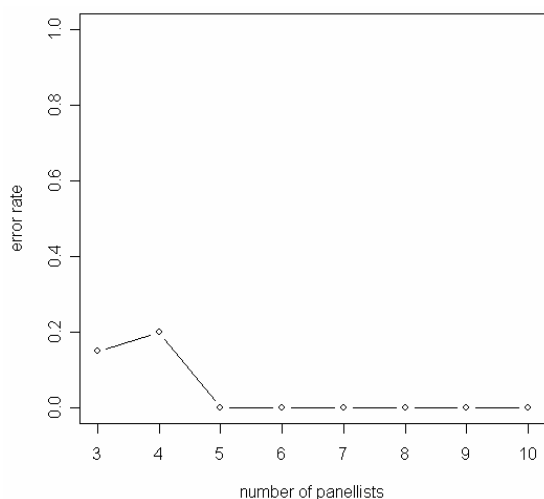
k) Error rate when 13 successes were observed. l) Error rate when 14 successes were observed.



m) Error rate when 15 successes were observed. n) Error rate when 16 successes were observed.



o) Error rate when 17 successes were observed. p) Error rate when 18 successes were observed.



r) Error rate when 19 successes were observed.

Figure 6.3: Collection of line plots indicating how the error rates vary for a fixed number of successes for various panel sizes.

Table 6.3 contains the error rates for the distinct panel sizes for a fixed number of successes (x). When 11 to 17 successes are recorded, the accuracy of small panels is extremely low. This can be attributed to the fact that the significance level for the complete panel is approximately 5% which was the required significance level for the triangle test. Thus for small panels, a single panellist could cause a false insignificant or significant result and therefore the error rate is very high. A panel of at least 9 is consequently recommended to obtain relatively reliable results.

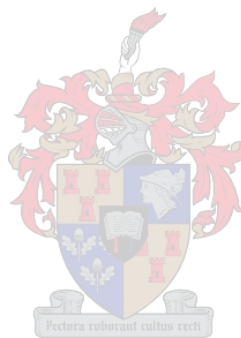
Table 6.3: Summary of the different error rates for various panel sizes and number of successes.

	Panel size								
x	3	4	5	6	7	8	9	10	Signif. level
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9997
2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9967
3	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9824
4	0.0035	0.0002	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000	0.9396
5	0.0088	0.0010	0.0049	0.0004	0.0014	0.0000	0.0000	0.0000	0.8485
6	0.0175	0.0031	0.0139	0.0022	0.0072	0.0007	0.0022	0.0000	0.7028
7	0.0307	0.0072	0.0307	0.0072	0.0223	0.0044	0.0124	0.0015	0.5207
8	0.0491	0.0144	0.0578	0.0181	0.0521	0.0154	0.0399	0.0099	0.3385
9	0.0737	0.0260	0.0975	0.0379	0.1018	0.0399	0.0949	0.0349	0.1905
10	0.1053	0.0433	0.1517	0.0704	0.1749	0.0849	0.1849	0.0894	0.0919
11	0.8553	0.9319	0.7786	0.8808	0.7276	0.8431	0.6890	0.8151	0.0376
12	0.8070	0.8978	0.6935	0.8127	0.6084	0.7404	0.5350	0.6750	0.0130
13	0.7491	0.8524	0.5942	0.7233	0.4743	0.6084	0.3700	0.5000	0.0037
14	0.6807	0.7934	0.4835	0.6126	0.3359	0.4551	0.2167	0.3142	0.0009
15	0.6009	0.7183	0.3661	0.4835	0.2068	0.2962	0.0975	0.1517	0.0002
16	0.5088	0.6244	0.2487	0.3426	0.1011	0.1531	0.0260	0.0433	0.0000
17	0.4035	0.5088	0.1404	0.2018	0.0307	0.0491	0.0000	0.0000	0.0000
18	0.2842	0.3684	0.0526	0.0789	0.0000	0.0000	0.0000	0.0000	0.0000
19	0.1500	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

To conclude...

Some characteristics of the paired comparison as well as triangle tests were considered in this section. It was discovered that for specific variables the results of the biplot correspond to those obtained in a paired comparison trial. The biplot was however implemented as an exploratory device and it is not recommended to abandon paired comparison tests to simply apply biplot methodology. Further study is necessary before such conclusions can be made.

The influence of the panel size in triangle tests was also of interest. As expected, a smaller panel may be sufficient in trivial trial cases, i.e. where there is no difference at all or there is an obvious difference between the samples. In trials where the significance level is borderline, a single individual may be extremely influential. Consequently, a panel of at least 9 is recommended, but to obtain truly consistent results, even more panellists may be needed.



CHAPTER 7

CONCLUSION

Sensory science enjoys a wide variety of applications in practice. The procedures available to analyse questions in this field are well developed and thorough descriptions of the methodology as well as corresponding assumptions necessary to implement them, are available.

It is also found in practice that these procedures are typically applied to the field of product development and are implemented by researchers specialising in psychology, chemistry or sensory sciences with a limited statistical background. These individuals are often more interested in the physical composition and characteristics of a product than the sensory aspects corresponding to them. Thus the sensory procedures are often carried out without much consideration as to whether they are appropriate for the hypothesis being examined. Furthermore, the methodology necessary to carry out a trial as well as the relevant corresponding critical value can also be prescribed by an organisation, while not supplying any information concerning the assumptions that need to be satisfied.

Sensory trials, performed at a specific product development centre, were considered in this study. Several different trials were assessed and the distributions of the data and test statistics received some attention. Table 7.1 contains a summary of the results obtained when implementing certain parametric and non-parametric test procedures as well as the organisation's IMQM test procedure to these data sets. The parametric tests consisted of t-tests as well as significance levels obtained by assuming a truncated absolute normal distribution. These are followed by non-parametric bootstrap and permutation tests, also included in Table 7.1.

Table 7.1: Summary of the results (significance levels and ASLs) obtained for all the parametric and non-parametric test procedures.

		Parametric test procedures (significance values shown)			Non-parametric test procedures (ASL-values shown)					
Data set	Null Hypothesis	Parametric t-test		Truncated Absolute Normal Distribution	Bootstrap test					Permutation test
		One-sample	Two-sample		One-sample	Two-sample	IMQM one	IMQM mean	IMQM 95-percentile	
3-day data	$H_0 : \mu_{C3} = 0$	0.0477		0.0455	0.0200					
	$H_0 : \mu_{T3} = 0$	0.0016		0.4491	0.0030		0.9820	0.5370	0.0805	
	$H_{01} : \mu_{T3} = \mu_{C3}$	0.0045	0.0019		0.0060	0.0005				0.0005
9-day data	$H_0 : \mu_{C9} = 0$	0.0519		0.2207	0.0210					
	$H_0 : \mu_{T9} = 0$	0.0011		0.4017	0.0050		0.8430	0.5830	0.0565	
	$H_{04} : \mu_{T9} = \mu_{C9}$	0.0216	0.0255		0.0330	0.0155				0.0155
Combined data	$H_{02} : \mu_{T9} = \mu_{C3}$		0.0042			0.0025				0.0025
	$H_{03} : \mu_{C9} = \mu_{C3}$		0.2823			0.4715				0.4715
	$H_{05} : \mu_{T3} = \mu_{C9}$		0.0064			0.0030				0.0030
	$H_{06} : \mu_{T9} = \mu_{T3}$		0.8856			0.1580				0.1580
	$H_0 : \mu_C = 0$	0.0093		0.1397	0.0010					

The aim of the taint tests performed in the 3-day and 9-day trials is to determine whether placing consumable products in freshly printed product packaging can cause a taint in the taste of the product. If a taint was perceived, the question to be addressed became whether “airing” the printed packaging for a longer period, for instance nine days, before placing it near the product would prevent the products from attaining a taint. This was tested by the IMQM procedure which ignores the scores allocated to the coded control items. The significance level obtained by fitting a truncated absolute normal distribution to either set of allocated test scores revealed that the significance levels used in the two tests are very high (0.4491 and 0.4017, respectively) when one is used as critical value.

The IMQM test and the significance level obtained by the fitted truncated normal distribution would lead to different conclusions. Even when testing the scores allocated to the coded control samples, significance levels of 0.0455, 0.2207 and 0.1397 were obtained for the critical value of one. Since it is known that the coded control samples come from the null distribution, the latter estimates may give a better representation of the true significance levels. Only the 0.0455 value is satisfactory and thus it appears that the probability of a type I error in the IMQM procedure may be ominously large. The significance levels also vary substantially, which raises the concern about the choice of the mean as test statistic. The mean does not account for the variation in the data, which may have a large influence on the obtained significance level of a test.

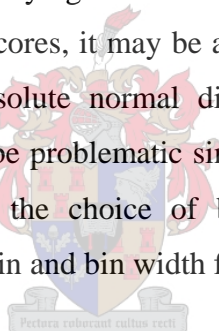
The parametric t-test results for testing whether the mean of the control data vary significantly from zero, differ from those obtained in the IMQM as well as from the truncated absolute normal test. Here it seems that the IMQM and truncated absolute normal test appear to be more trustworthy. In these results for the test data however, the IMQM and t-test rendered similar results but the results from the truncated absolute normal test were different.

Both the permutation and bootstrap tests render similar results in two-sample cases, although the ASLs in the permutation tests are lower than those in the bootstrap tests. The IMQM results do not correspond to those in the bootstrap tests for the control data, but

they do correspond to those of the test data. Still, the IMQM method is not recommended since it makes no adjustment for the variability in the data.

Due to the small sample size and the fact that only two trials' data are available, no concrete alternative critical value to implement in the IMQM procedure can be recommended at this stage. More trials consisting of larger samples are necessary to estimate the appropriate critical value. However, the following preliminary method is suggested that does take the variability of the data into account. Use the test scores to fit an appropriate truncated normal distribution under the null hypothesis as was done in Chapter 3. Compute the $(1-\alpha)$ -quantile for this estimated distribution. This quantile is then employed as a critical value which is compared to the mean of the test scores.

Since the observations from the underlying continuous truncated distribution function are measured in a discrete way by the scores, it may be appropriate to apply a discretisation process to the fitted truncated absolute normal distribution prior to estimating the appropriate critical value. This can be problematic since, similar to histograms; the final fitted distribution will depend on the choice of bin origin as well as bin width. Determining the appropriate bin origin and bin width fall beyond the scope of this study.



The choice of panel size in triangle as well as taint tests received some attention. To be able to prescribe the specific number of minimum panellists required, a more intensive study is needed. It can be concluded however that panels of seven or less will rarely be adequate to obtain reliable and valid results.

Biplots proved to be very informative when considering sensory trials surrounding the modification of products. It is suggested that this technique be used as an exploratory tool in conjunction with sensory trials, as it gives a multidimensional perspective on the effect of changes in the composition of a product on the perceived taste of that product.

In this study only analytical sensory trials were considered. When, however, considering hedonic questions, another multivariate technique that can be useful in preference testing

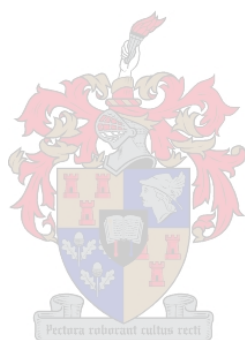
is unfolding. Borg & Groenen (1997) states that unfolding “assumes that different individuals perceive various objects of choice in the same way but differ with respect to what they consider an ideal combination of the objects’ attributes.” Unfolding is a multidimensional scaling procedure that strives to represent panelists and objects in such a way that the distance from each panelist to every object mirrors their preference.

In this assignment the several sensory procedures, carried out at a specific product development centre, were considered. The following has been shown:

- The IMQM procedure does not seem to yield satisfactory significance levels.
- The specified number of panellists required is not sufficient.
- The power of the sensory tests is inadequate and needs some attention.

Since the data available were limited, only preliminary solutions to these issues can be offered at this stage. If, however, the organisation intends to continue using these sensory procedures, it is recommended that more data should be obtained to gain more clear cut solutions.





REFERENCES

- ALPERN, M., LAWRENCE, M. & WOLSK, D. (1967). *Sensory Processes*. Brooks-Cole, Belmont, California.
- BORG, I. & GROENEN, P.J.F. (2002). *Modern Multidimensional Scaling: Theory and Applications, Second edition*. Springer-Verlag, New York.
- BOWER, J.A. (1996). Statistics for food science III: sensory evaluation data. Part B – discrimination tests. *Nutrition & Food Science*, **2**, 16–22.
- BRITISH STANDARDS INSTITUTION, 1982. *BS 5929: Methods for Sensory Analysis of Foods Part 2: Paired Comparison Test*, British Standards Institution, London.
- BROCKHOFF, P.M.B. (1995). Generalised linear models in sensometrics. In *Proceedings: 4th European Conference on Food Industry and Statistics* (pp. 33–47), 7–8 December, Dijon, France.
- BROCKHOFF, P.B. (2003). The statistical power of replications in difference tests. *Food Quality and Preference*, **14**, 405–417.
- BROCKHOFF, P.B. & SCHLICH, P. (1998). Handling replications in discrimination tests. *Food Quality and Preference*, **9**, 303–312.
- BYRD, R.H., LU, P., NOCEDAL, J. & ZHU, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, **16**, 1190–1208.
- COX, T.F. & COX, M.A.A (2001). *Multidimensional Scaling, Second edition*. Chapman & Hall/CRC, London.
- DIJKSTERHUIS, G.B. (1997). *Multivariate data analysis in sensory and consumer science*. Food & Nutrition Press Inc., Trumbull, Connecticut.
- EFRON, B. & TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, New York.
- ENNIS, D.M. & BI, J. (1998). The beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies*, **13**, 3819–4112.

- GOOD, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Second Edition*. Springer-Verlag, New York.
- GOWER, J.C. & HAND, D.J. (1996). *Biplots*. Chapman & Hall, London.
- HUNTER, E.A., PIGGOT, J.R. & LEE, K.Y.M. (2000). Analysis of discrimination tests. In *Proceedings: 6th Conference on Food Industry and Statistics* (pp. 9–98), 19–21 January, Pau, France.
- MOOD, A.M., GRAYBILL, F.A., & BOES, D.C. (1963). *Introduction to the theory of statistics*. McGraw-Hill, New York.
- OSGOOD, C.E. (1953). *Method and Theory in Experimental Psychology*. Oxford University Press, New York.
- SCOTT, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons Inc., New York.
- SCHLICH, P. (1993). Risk tables for discrimination tests. *Food Quality and Preference*, **4**, 141–151.
- SIEGEL, S. (1956). *Nonparametric statistics for the behavioural sciences*. McGraw-Hill, New York.
- STONE, H. & SIDEL, J.L. (1985). *Sensory evaluation practices*, First edition. Academic Press, Inc., London.
- WACKERLY, D.D., MENDENHALL, W. & SCHEAFFER, R.L. (2002). *Mathematical Statistics with Applications*. Duxbury, Pacific Grove.
- WOODWORTH, R.S. & SCHLOSBERG, H. (1960). *Experimental psychology*. Holt, New York.
- ZAR, J.H. (1996). *Biostatistical Analysis, Third Edition*. Prentice Hall, Upper Saddle River, New Jersey.

APPENDIX

The appendix contains the following:

- 14.1 R-function `MLE.estimate()` implemented in Section 3.4.
- 14.2 R-function `shaded.area.graph()` implemented in Section 3.4
- 14.3 R-function `crit.val.trunc.norm()` implemented in Section 3.4
- 14.4 R-function `repeat.taint.test()` implemented in Section 3.5
- 14.5 R-function `bootstrap.dist()` implemented in Section 4.2
- 14.6 R-function `bootstrap.dens()` implemented in Section 4.2.5
- 14.7 R-function `boot.test()` implemented in Section 4.3
- 14.8 R-function `one.smpl.boot.test()` implemented in Section 4.3
- 14.9 R-function `boot.test.bipl()` implemented in Section 4.4
- 14.10 R-function `PCAbipl()` implemented in Section 4.4
- 14.11 R-function `blegend()` implemented in Section 4.4
- 14.12 R-function `drawbipl()` implemented in Section 4.4
- 14.13 R-function `permutation.test()` implemented in Section 4.4
- 14.14 R-function `real.perm.dist()` implemented in Section 4.4
- 14.15 R-function `accuracy.perm.test()` implemented in Section 4.5
- 14.16 R-function `repeat.triangle.test()` implemented in Section 4.5

14.1 R-function `MLE.estimate()`

```
MLE.estimate
function (datvec, graph=TRUE, mtd="Nelder-Mead", lwr=-Inf, ppr=Inf,
x.lab="test scores")
{
#function performing numerical maximisation of the log of the
#likelihood function for the parameters
#of a normal distribution truncated at 0 and 5
#datvec is a vector of values used to fit the truncated distribution

mu.begin<-mean(datvec)
sigma.begin<-var(datvec)
```

```

mle<-optim(par=c(mu.begin,sigma.begin),fn=function(x){a<-x[1];b<-
x[2];- sum(log(
dabsnorm(datvec,a,b)/pabsnorm(5,a,b)))},method=mtd,lower=lwr,upper=ppr
)
mle.null<-optim(par=c(sigma.begin),fn=function(x){b<-x;- sum(log(
dabsnorm(datvec,0,b)/pabsnorm(5,0,b)
))},method=mtd,lower=lwr,upper=ppr)

par.alternative<-mle$par
MLE.alternative<-mle$value

sigma.null<-mle.null$par
MLE.null<-mle.null$value

x<-seq(from=0,to=5,len=1000)
pdf.alternative<-
dabsnorm(x,par.alternative[1],par.alternative[2])/(pabsnorm(5,par.alte
rnative[1],par.alternative[2]))
pdf.null<-dabsnorm(x,0,sigma.null)/(pabsnorm(5,0,sigma.null))
y.lim<-c(0,.4+max(c(pdf.alternative,pdf.null)))

if(graph){
hist(datvec,freq=FALSE,breaks=seq(from=-
.25,to=5.25,by=.5),main="alternative
hypothesis",xlab=x.lab,ylab="density",xlim=c(0,5),ylim=y.lim)
lines(x,pdf.alternative,col="red")
win.graph()

hist(datvec,freq=FALSE,breaks=seq(from=-.25,to=5.25,by=.5),main="null
hypothesis",xlab=x.lab,ylab="density",xlim=c(0,5),ylim=y.lim)
lines(x,pdf.null,col="blue")
win.graph()

hist(datvec,freq=FALSE,breaks=seq(from=-
.25,to=5.25,by=.5),main="",xlab=x.lab,ylab="density",xlim=c(0,5),ylim=
y.lim)
lines(x,pdf.alternative,col="red")
lines(x,pdf.null,col="blue")
}

```

```
list(estimates.under.Ha=par.alternative,MLE.Ha=MLE.alternative,estimates.under.H0=sigma.null,MLE.H0=MLE.null)
}
```

14.2 R-function shaded.area.graph()

```
shaded.area.graph
function (crit.val=1,mu,sigma,num.lines=350,y.lim=c(0,.5),upper=TRUE)
{
#function used to create probability density graph
#for specific estimates of mu and sigma
#shading that graph above or below a given critical value
#calculating the shaded area

pdf.func<-function(x) dabsnorm(x,mu,sigma)/(pabsnorm(5,mu,sigma)-
pabsnorm(0,mu,sigma))
x<-seq(from=0,to=5,length.out=1000)
pdf<-pdf.func(x)

plot(x,pdf,type="l", ylim=y.lim)
abline(v=crit.val)
abline(h=0)
abline(v=0)
ifelse(upper,x.seq<-
seq(from=crit.val,to=5,length.out=num.lines),x.seq<-
seq(from=0,to=crit.val,length.out=num.lines))
sapply(x.seq, function(x) lines(matrix(c(x,x,0,pdf.func(x)),nrow=2)))
ifelse(upper,shaded.area<-1-
((pabsnorm(crit.val,mu,sigma))/(pabsnorm(5,mu,sigma))),shaded.area<-
((pabsnorm(crit.val,mu,sigma))/(pabsnorm(5,mu,sigma))))

list(shaded.area=shaded.area)
}
```

14.3 R-function crit.val.trunc.abs.norm()

```
crit.val.trunc.abs.norm
function (alfa=0.05,mu,sigma2,num.iter=3000,stop.krit=0.00001)
```

```

{
#function to find the (1-alfa)th percentile
#for truncated absolute normal distribution
#with parameters mu and sigma2
x.vec<-seq(from=0,to=5,length.out=num.iter)
y.vec<-pabsnorm(x.vec,mu=mu,sigma=sigma2)
p<-(1-alfa)*pabsnorm(5,mu=mu,sigma=sigma2)
x.close<-x.vec[round(y.vec,2)==round(p,2)]
y.close<-y.vec[round(y.vec,2)==round(p,2)]
x<-x.close[ which.min(abs(y.close-p))]
p.krit<-pabsnorm(x,mu=mu,sigma=sigma2)
for(i in 1:num.iter)
{
if( abs(p.krit-1+alfa) > stop.krit)
{
x.vec<-seq(from=x-.1,to=x+.1,length.out=num.iter)
y.vec<-pabsnorm(x.vec,mu=mu,sigma=sigma2)
x.close<-x.vec[round(y.vec,2)==round(p,2)]
y.close<-y.vec[round(y.vec,2)==round(p,2)]
x<-x.close[ which.min(abs(y.close-p))]
p.krit<-pabsnorm(x,mu=mu,sigma=sigma2)
}
else{break}
}
list(mu=mu,sigma2=sigma2,x.crit=x,associat.alfa=p.krit/pabsnorm(5,mu,sigma2))
}

```

14.4 R-function repeat.taint.test()

```

repeat.taint.test
function (z.vec,num.panl=3)
{
library(combinat)
poss.comb<-combn(length(z.vec),num.panl)

mean.vec<-apply(poss.comb,2,function(x) mean(z.vec[x]))
n<-length(mean.vec)
result.func<-function(x) ifelse(x>1,"reject null hypothesis","cannot
reject null hypothesis")

```

```

true.mean<-mean(z.vec)
true.result<-result.func(true.mean)
error.rate<-ifelse(true.mean>1,sum(mean.vec<=1)/n,sum(mean.vec>1)/n)
list(mean.vec=mean.vec,true.result=true.result,error.rate=error.rate)
}

```

14.5 R-function bootstrap.dist()

```

bootstrap.dist
function (data,B=1000,control=F)
#data is a matrix where the first column contain the control values
#and the second column the test values
{
n<-nrow(data)
mu.vec<-c()
for (i in 1:B) {
stkprf<-sample(1:n,n,replace=TRUE)
if(control==F){
boot.stkprf<-data[stkprf,1]
mu<-mean(data[,1])
std.err<-(var(data[,1]))^.5/(sqrt(n))}
else{boot.stkprf<-data[stkprf,2]
mu<-mean(data[,2])
std.err<-(var(data[,2]))^.5/(sqrt(n))}
}
mu.vec[i]<-mean(boot.stkprf)
}
boot.std.err<-(var(mu.vec))^.5
hist(mu.vec,main="Histogram of bootstrap
mean",xlab="mean",freq=FALSE,breaks=25)
abline(v=mu,lty=2)
list(mu=mu,std.err=std.err,boot.std.err=boot.std.err)
}

```



14.6 R-function bootstrap.dens()

```

bootstrap.dens
function (data,B=4000,control=F,bw,x.limit=c(-.3,1))
#data is a matrix where the first column contain the control values
#and the second column the test values
{

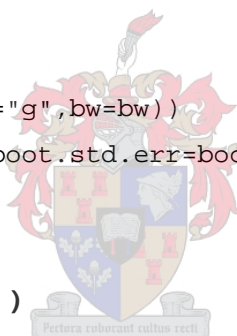
```

```

n<-nrow(data)
mu.vec<-c()
for (i in 1:B) {
  stkprf<-sample(1:n,n,replace=TRUE)
  if(control==F){
    boot.stkprf<-data[stkprf,1]
    mu<-mean(data[,1])
    std.err<-(var(data[,1]))^.5/(sqrt(n))}
  else{boot.stkprf<-data[stkprf,2]
    mu<-mean(data[,2])
    std.err<-(var(data[,2]))^.5/(sqrt(n))
  }
  mu.vec[i]<-mean(boot.stkprf)
}
boot.std.err<-(var(mu.vec))^.5
hist(mu.vec,main="",xlab="mean",freq=FALSE,breaks=25,xlim=x.limit)
abline(v=mu,lty=2)
abline(v=0,lty=1,col="red")
lines(density(mu.vec,kernel="g",bw=bw))
list(mu=mu,std.err=std.err,boot.std.err=boot.std.err)
}

```

14.7 R-function boot.test()



```

boot.test
function
(z.vec=craven3dae[,1],y.vec=craven3dae[,2],B=2000,y.lim=c(0,B/2),one.s
ided=TRUE)
{
  n<-length(z.vec)
  one.smpl.t.test.stat.vec<-c()
  two.smpl.t.test.stat.vec<-c()
  std.test.vec<-c()
  x.vec<-z.vec-y.vec
  x.star<-x.vec-mean(x.vec)
  vec<-c(z.vec,y.vec)
  sigma<-sd(x.vec)/sqrt(n)
  sigma.z<-var(z.vec)/n
  sigma.y<-var(y.vec)/n
  one.smpl.t.test.stat<-(mean(x.vec))/sigma

```

```

two.smpl.t.test.stat<-(mean(z.vec)-mean(y.vec))/sqrt(sigma.z+sigma.y)
std.test<-mean(z.vec)
param.signif.one.smpl<-1-pt(one.smpl.t.test.stat,(n-1))
param.signif.two.smpl<-1-pt(two.smpl.t.test.stat,(2*n-2))
for(i in 1:B) {

    smpl<-sample(1:n,n,replace=TRUE)
    two.smpl<-sample(1:(2*n),2*n,replace=TRUE)
    boot.z<-vec[two.smpl[1:n]]
    boot.y<-vec[two.smpl[(n+1):(2*n)]]
    sigma.x<-sd(x.star[smpl])/sqrt(n)
    sigma.z.boot<-var(boot.z)/n
    sigma.y.boot<-var(boot.y)/n
    one.smpl.t.test.stat.vec[i]<-(mean(x.star[smpl]))/sigma.x
    two.smpl.t.test.stat.vec[i]<-(mean(boot.z)-
mean(boot.y))/sqrt(sigma.z.boot+sigma.y.boot)
    std.test.vec[i]<-mean(z.vec[smpl])
}
one.smpl.t.test.stat.vec<-
one.smpl.t.test.stat.vec[!is.nan(one.smpl.t.test.stat.vec)]
two.smpl.t.test.stat.vec<-
two.smpl.t.test.stat.vec[!is.nan(two.smpl.t.test.stat.vec)]
if(one.sided){
ASL.one.smpl.t.test<-
sum(one.smpl.t.test.stat.vec>=one.smpl.t.test.stat)/length(one.smpl.t.
test.stat.vec)
ASL.two.smpl.t.test<-
sum(two.smpl.t.test.stat.vec>=two.smpl.t.test.stat)/length(two.smpl.t.
test.stat.vec)}

else{
ASL.one.smpl.t.test<-
sum(one.smpl.t.test.stat.vec>=one.smpl.t.test.stat)/length(one.smpl.t.
test.stat.vec)
ASL.two.smpl.t.test<-
sum(two.smpl.t.test.stat.vec>=two.smpl.t.test.stat)/length(two.smpl.t.
test.stat.vec)}
ASL.QM.crit.val.1<-sum(std.test.vec>=1)/B
ASL.QM.crit.val.mean<-sum(std.test.vec>=std.test)/B

```



```

optimal.crit.val<-quantile(std.test.vec,.95)
ASL.MQM.optimal.crit.val<-sum(std.test.vec>=optimal.crit.val)/B
result.boot.one.smpl.t<-ifelse(ASL.one.smpl.t.test<=0.05,"null
hypothesis is rejected at 5% significance level","can not reject null
hypothesis")
result.boot.two.smpl.t<-ifelse(ASL.two.smpl.t.test<=0.05,"null
hypothesis is rejected at 5% significance level","can not reject null
hypothesis")
result.boot.MQM.crit.val.1<-ifelse(ASL.MQM.crit.val.1<=0.05,"null
hypothesis is rejected at 5% significance level","can not reject null
hypothesis")
result.boot.MQM.crit.val.mean<-ifelse(ASL.MQM.crit.val.1<=0.05,"null
hypothesis is rejected at 5% significance level","can not reject null
hypothesis")
result.boot.MQM.optimal.crit.val<-
ifelse(ASL.MQM.optimal.crit.val<=0.05,"null hypothesis is rejected at
5% significance level","can not reject null hypothesis")
hist(one.smpl.t.test.stat.vec,xlab="theta",main="one-sample t-
test",ylim=y.lim)
abline(v=one.smpl.t.test.stat, lty=2)
win.graph()
hist(two.smpl.t.test.stat.vec,xlab="theta",main="two-sample t-
test",ylim=y.lim)
abline(v=two.smpl.t.test.stat, lty=2)
win.graph()
hist(std.test.vec,xlab="theta",main="std.procedure",ylim=y.lim)
abline(v=std.test,lty=2)
abline(v=1,lty=3,col="red")
abline(v=optimal.crit.val,col="green")

list(one.smpl.t.test.stat=one.smpl.t.test.stat,param.signif.one.smpl=p
aram.signif.one.smpl,
ASL.one.smpl.t.test=ASL.one.smpl.t.test,two.smpl.t.test.stat=two.smpl.
t.test.stat,param.signif.two.smpl=1-pt(two.smpl.t.test.stat,(2*n-2)),
ASL.two.smpl.t.test=ASL.two.smpl.t.test,std.test=std.test,ASL.MQM.crit
.val.1=ASL.MQM.crit.val.1,
result.boot.MQM.crit.val.1=result.boot.MQM.crit.val.1,ASL.MQM.crit.val
.mean=ASL.MQM.crit.val.mean,
result.boot.MQM.crit.val.mean=result.boot.MQM.crit.val.mean,

```

```

optimal.crit.val=optimal.crit.val,ASL.MQM.optimal.crit.val=ASL.MQM.opt
imal.crit.val,
result.boot.MQM.optimal.crit.val=result.boot.MQM.optimal.crit.val)
}

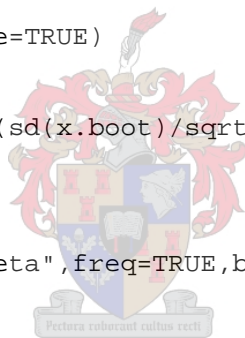
```

14.8 R-function `one.smpl.boot.test()`

```

one.smpl.boot.test
function (z.vec=craven9dae[,2],y.vec=craven3dae[,2],B=2000)
{
  theta.vec<-c()
  aantal<-c()
  x<-c(z.vec,y.vec)
  theta<-mean(x)/(sd(x)/sqrt(length(x)))
  x.bar<-x-mean(x)
  n<-length(x)
  for( i in 1:B){
    x.smpl<-sample(1:n,n,replace=TRUE)
    x.boot<-x.bar[x.smpl]
    theta.vec[i]<-mean(x.boot)/(sd(x.boot)/sqrt(n))
  }
  ASL<-mean(theta.vec>theta)
  hist(theta,main="",xlab="theta",freq=TRUE,breaks=25)
  abline(v=theta,lty=2)
  list(theta=theta,param.signif.t.test=1-pt(theta,n-
1),ASL.one.sample=ASL)
}

```



14.9 R-function `boot.test.bipl()`

```

boot.test.bipl
function (z.vec,y.vec,B=2000,y.lim=c(0,B),x.lim=c(-3,3))
{
  library(MASS)

  vec<-c(z.vec,y.vec)
  N<-length(vec)
  n<-length(z.vec)

  stats.mat<-matrix(ncol=10,nrow=B)

```

```

stats.func<-function(z,y) c(mean(z)-mean(y),
quantile(z,probs=seq(from=.1,to=.9,by=.1))-
quantile(y,probs=seq(from=.1,to=.9,by=.1))      )

bootmat<-matrix(sample(1:N,N*B,replace=TRUE),nrow=N,ncol=B)
for ( i in 1:B)
  {
    boot.z.vec<-(vec[bootmat[,i]][1:n]
    boot.y.vec<-(vec[bootmat[,i]][(n+1):N]

stats.mat[i,]<-stats.func(boot.z.vec,boot.y.vec)
  }
colnames(stats.mat)<-c("mean",paste(seq(from=10,to=90,by=10),"th
percentile",sep=""))

for( j in 1:ncol(stats.mat))
  {win.graph()

hist(stats.mat[,j],xlab=colnames(stats.mat)[j],main="",ylim=y.lim,xlim
=x.lim)
  }
win.graph()
PCAbipl(stats.mat)
}

```



14.10 R-function PCAbipl()

```

PCAbipl
function (X, Y = NULL, X.new = NULL, scaled.mat = T, e.vects =
1:ncol(X),
  ax = 1:ncol(X), plotchr = 15, label = T, markers = T, Title =
NULL,
  n.int = rep(5, ncol(X)), offset = rep(0, ncol(X)))
{
  colours <- c(8, 4, 6, 3, 7, 15, 1, 2, 5, 9, 11, 13, 10, 14,
    16, 12)
  unscaled.X <- X
  means <- apply(X, 2, mean)
  sd <- sqrt(apply(X, 2, var))

```

```

if (scaled.mat)
  X <- scale(X)
else X <- scale(X, scale = F)
n <- nrow(X)
p <- ncol(X)
if (is.null(Y))
  J <- 0
else J <- ncol(Y)
while (J > length(colours)) colours <- c(colours, colours)
if (is.null(dimnames(X)))
  dimnames(X) <- list(paste(1:n), paste("V", 1:p, sep = ""))
if (length(dimnames(X)[[1]]) == 0)
  dimnames(X)[[1]] <- paste(1:n)
if (length(dimnames(X)[[2]]) == 0)
  dimnames(X)[[2]] <- paste("V", 1:p, sep = "")
if (J > 0) {
  if (nrow(Y) != n)
    stop("number of rows of X and Y differ")
  if (is.null(dimnames(Y)))
    dimnames(Y) <- list(NULL, paste("class", 1:J, sep = ""))
  if (length(dimnames(Y)[[2]]) == 0)
    dimnames(Y)[[2]] <- paste("class", 1:J, sep = "")
}
Vr <- svd(t(X) %*% X)$u[, e.vects[1:2]]
eigval <- svd(t(X) %*% X)$d
Z <- X %*% Vr
Z <- cbind(Z, plotchr)
Z <- cbind(Z, 1)
if (J > 0)
  for (j in 1:J) Z[Y[, j] == 1, 4] <- colours[j]
axes.rows <- 1/(diag(Vr %*% t(Vr))) * Vr
z.axes <- lapply(1:p, function(j, unscaled.X, means, sd,
  axes.rows, n.int) {
  number.points <- 30
  std.markers <- pretty(unscaled.X[, j], n = n.int[j])
  std.range <- c(min(std.markers), max(std.markers))
  std.markers.min <- std.markers - (std.range[2] - std.range[1])
  std.markers.max <- std.markers + (std.range[2] - std.range[1])

```

```

      std.markers <- c(std.markers, std.markers.min,
std.markers.max)
      interval <- (std.markers - means[j])/sd[j]
      axis.vals <- seq(from = min(interval), to = max(interval),
        length = number.points)
      axis.vals <- sort(unique(c(axis.vals, interval)))
      number.points <- length(axis.vals)
      axis.points <- matrix(0, nrow = number.points, ncol = 4)
      axis.points[, 1] <- axis.vals * axes.rows[j, 1]
      axis.points[, 2] <- axis.vals * axes.rows[j, 2]
      axis.points[, 3] <- axis.vals * sd[j] + means[j]
      axis.points[, 4] <- 0
      for (i in 1:number.points) if (any(zapsmall(axis.points[i,
        3] - std.markers) == 0))
        axis.points[i, 4] <- 1
      return(axis.points)
    }, unscaled.X = unscaled.X, means = means, sd = sd, axes.rows =
axes.rows,
      n.int = n.int)
    drawbipl.R(Z, z.axes, z.axes.names = dimnames(X)[[2]], ax = ax,
      label = label, markers = markers, offset = offset, Title =
Title)
    if (!(is.null(X.new))) {
      Z.new <- scale(X.new, means, sd) %*% Vr
      points(Z.new)
    }
    if (J > 0) {
      windows()
      blegend(dimnames(Y)[[2]], colours[1:J])
    }
    fit.quality <- (eigval[e.vects[1]] +
eigval[e.vects[2]])/sum(eigval)
    fit.adequacy <- diag(Vr %*% t(Vr))
    names(fit.adequacy) <- dimnames(X)[[2]]
    list(Eigenvectors = Vr, e.vals = eigval, quality = fit.quality,
      adequacy = fit.adequacy)
  }

```

14.11 R-function `blegend ()`

```
blegend
function (classes, colours)
{
  J <- length(colours)
  plot(x = c(0, 10), y = c(0, J + 1), type = "n", axes = F,
       xlab = "", ylab = "")
  for (j in 1:J) {
    points(x = 1, y = J - j + 1, pch = 18, cex = 2, col =
colours[j])
    text(x = 2, y = J - j + 1, classes[j], adj = 0)
  }
}
```

14.12 R-function `drawbipl ()`

```
drawbipl
function (Z, z.axes, z.axes.names = NULL, ax = NULL, exp.factor = 1.2,
        label = T, markers = T, Title = NULL, offset = rep(0,
length(z.axes)))
{
  par(pty = "s")
  p <- length(z.axes)
  eqscplot(Z[, 1] * exp.factor, Z[, 2] * exp.factor, xaxt = "n",
          yaxt = "n", xlab = "", ylab = "", type = "n")
  usr <- par("usr")
  if (ncol(Z) == 2)
    Z <- cbind(Z, 15)
  if (ncol(Z) == 3)
    Z <- cbind(Z, 1)
  if (ncol(Z) == 4)
    Z <- cbind(Z, 0.7)
  Z.plot <- Z
  x.vals <- Z.plot[, 1]
  y.vals <- Z.plot[, 2]
  invals <- x.vals < usr[2] & x.vals > usr[1] & y.vals < usr[4] &
    y.vals > usr[3]
  Z.plot <- Z.plot[invals, ]
```

```

    apply(Z.plot, 1, function(a) points(x = a[1], y = a[2], pch =
as.vector(a)[3],
      col = as.vector(a)[4], cex = as.vector(a)[5]))
  if (is.null(dimnames(Z)))
    dimnames(Z) <- list(paste("s", 1:nrow(Z), sep = ""),
      NULL)
  if (length(dimnames(Z)[[1]]) == 0)
    dimnames(Z) <- list(paste("s", 1:nrow(Z), sep = ""),
      dimnames(Z)[[2]])
  if (label == T)
    text(Z[, 1], Z[, 2] - 0.015 * (usr[4] - usr[3]), labels =
dimnames(Z)[[1]],
      cex = 0.65)
  if (is.null(ax))
    axes <- NULL
  else axes <- (1:p)[ax]
  for (i in axes) {
    marker.mat <- z.axes[[i]][z.axes[[i]][, 4] == 1, 1:3]
    x.vals <- marker.mat[, 1]
    y.vals <- marker.mat[, 2]
    if (y.vals[1] == y.vals[length(y.vals)])
      gradient <- 0
    else {
      if (x.vals[1] == x.vals[length(x.vals)])
        gradient <- Inf
      gradient <- (y.vals[1] -
y.vals[length(y.vals)])/(x.vals[1] -
      x.vals[length(x.vals)])
    }
    if (is.null(z.axes.names)) {
      axis.name <- paste("v", i, sep = "")
    }
    else {
      axis.name <- z.axes.names[i]
    }
    if (abs(gradient) < 1) {
      lines(c(usr[1], usr[2]), c(usr[1], usr[2]) * gradient)
      extreme <- marker.mat[nrow(marker.mat), ]
      if (extreme[1] < 0)

```

```

        mtext(axis.name, side = 2, line = 0.2, las = 2,
              at = usr[1] * gradient + offset, cex = 0.65)
      else mtext(axis.name, side = 4, line = 0.2, las = 2,
              at = usr[2] * gradient + offset, cex = 0.65)
    }
  else {
    lines(c(usr[3], usr[4])/gradient, c(usr[3], usr[4]))
    extreme <- marker.mat[nrow(marker.mat), ]
    if (extreme[2] < 0)
      mtext(axis.name, side = 1, line = -0.2, at =
usr[3]/gradient +
              offset, cex = 0.65)
    else mtext(axis.name, side = 3, line = 0.2, at =
usr[4]/gradient +
              offset, cex = 0.65)
  }
  invals <- x.vals < usr[2] & x.vals > usr[1] & y.vals <
usr[4] & y.vals > usr[3]
  x.invals <- x.vals[invals | c(invals[-1], F) | c(F, invals[-
length(invals)])]
  y.invals <- y.vals[invals | c(invals[-1], F) | c(F, invals[-
length(invals)])]
  std.markers <- marker.mat[, 3]
  x.invals <- x.vals[x.vals < usr[2] & x.vals > usr[1] &
y.vals < usr[4] & y.vals > usr[3]]
  y.invals <- y.vals[x.vals < usr[2] & x.vals > usr[1] &
y.vals < usr[4] & y.vals > usr[3]]
  tick.labels <- std.markers[x.vals < usr[2] & x.vals >
usr[1] & y.vals < usr[4] & y.vals > usr[3]]
  points(x.invals, y.invals, pch = 16, cex = 0.4)
  if (markers == T) {
    x.labvals <- x.invals
    y.labvals <- y.invals
  }
  else {
    x.labvals <- x.invals[c(1, length(x.invals))]
    y.labvals <- y.invals[c(1, length(y.invals))]
    tick.labels <- tick.labels[c(1, length(tick.labels))]
  }
}

```



```

      text(x.labvals, y.labvals - 0.018 * (usr[4] - usr[3]),
           labels = paste("", tick.labels, sep = "", "\n"),
           cex = 0.5)
    }
    title(main = (ifelse(is.null(Title), "Biplot", Title)))
  }

```

14.13 R-function `permutation.test()`

```

> permutation.test
function
(z.vec,y.vec,B=2000,one.sided=T,return.vec=FALSE,plot.hist=TRUE)
#z.vec is the test data
#y.vec is the control data
#use algorithm 15.1 from Efron & Tibshirani (1993)
{
  n<-length(z.vec)
  m<-length(y.vec)
  data.vec<-c(z.vec,y.vec)
  N<-n+m
  theta<-mean(z.vec)-mean(y.vec)
  theta.vec<-c()
  number<-c()

  for ( i in 1:B)
  {
    stkprf<-sample(1:N,n,replace=FALSE)
    boot.z.vec<-data.vec[stkprf]
    boot.y.vec<-data.vec[-stkprf]
    theta.vec[i]<-mean(boot.z.vec)-mean(boot.y.vec)

    if(
      if(one.sided){theta.vec[i]>theta}
      else (abs(theta.vec[i])>abs(theta))
    )

    {number[i]<-1 }
    else{number[i]<-0}
  }
  approx.ASL<-mean(number)

```



```

if(approx.ASL<0.01) {
conclusion<-"very strong evidence against H0"}
if(approx.ASL>=0.01 && approx.ASL<0.025) {
conclusion<-"strong evidence against H0"}
if(approx.ASL>=0.025 && approx.ASL<0.05) {
conclusion<-"reasonably strong evidence against H0"}
if(approx.ASL>=0.05 && approx.ASL<0.1) {
conclusion<-"borderline evidence against H0"}
if(approx.ASL>=0.1) {
conclusion<-"cannot reject H0"}
if(plot.hist){
hist(theta.vec,main="Permutation distribution of the
mean",xlab="mean",freq=FALSE,breaks=25)
abline(v=theta,lty=2)}
if(return.vec)
{list(theta.vec=theta.vec,theta=theta,approx.ASL=cbind(approx.ASL,conc
lusion))}
else{
list(theta=theta,approx.ASL=cbind(approx.ASL,conclusion))}
}

```

14.14 R-function `real.perm.dist()`

```

real.perm.dist
function (z.vec,y.vec,list.mean.vec=FALSE)
{
library(combinat)
n<-length(z.vec)
m<-length(y.vec)
N<-n+m
vec<-c(z.vec,y.vec)
mean.vec<-c()
mat<-combn(N,n)
for ( i in 1:ncol(mat))
{z.vect<-vec[mat[,i]]
y.vect<-vec[-mat[,i]]
mean.vec[i]<-mean(z.vect)-mean(y.vect)
}
hist(mean.vec,xlab="mean",main="",breaks=25,freq=FALSE)

```

```

if(list.mean.vec){list(mean.vec=mean.vec,uniq.mean=sort(unique(mean.vec))
),numb.uniq.mean=length(sort(unique(mean.vec))))
}
else{
list(uniq.mean=sort(unique(mean.vec)),numb.uniq.mean=length(sort(unique
(mean.vec))))
}}

```

14.15 R-function `accuracy.perm.test()`

```

accuracy.perm.test
function (z.vec,y.vec,perm.smpl=2000,boot.smpl=1000)
{
#z.vec is the test data
#y.vec is the control data
approx.ASL.vec<-c()
theta<-mean(z.vec)-mean(y.vec)
theta.vec<-c()
diff.theta.vec<-c()
perm<-real.perm.dist(z.vec=z.vec,y.vec=y.vec,list.mean.vec=TRUE)
perm.vec<-perm[[1]]
diff.theta<-perm$numb.uniq.mean
ASL<-mean(perm.vec>theta)
for(i in 1:boot.smpl){

approx.perm<-
permutation.test(z.vec=z.vec,y.vec=y.vec,return.vec=TRUE,B=perm.smpl,p
lot.hist=FALSE)
approx.ASL.vec[i]<-mean(approx.perm[[1]]>theta)
diff.theta.vec[i]<-length(unique(approx.perm[[1]]))
}
hist(approx.ASL.vec,xlab="approximated ASL", main="")
abline(v=ASL,lty=2)
list(ASL=ASL,diff.theta=diff.theta,avg.diff.theta=mean(diff.theta.vec)
,sd.err.approx.ASL=sd(approx.ASL.vec),mean.ASL.vec=mean(approx.ASL.vec
))
}

```

14.16 R-function repeat.triangle.test()

```
> repeat.triangle.test
function (N=20,num.corr=1:19,num.panl=3:10,signif=0.05)
{
  library(combinat)
  error.mat<-matrix(ncol=length(num.panl),nrow=length(num.corr))
  rownames(error.mat)<-num.corr
  colnames(error.mat)<-num.panl
  true.signif<-c()
  true.result<-c()
  for(j in 1:length(num.corr)){
    vec<-c(rep(1,num.corr[j]),rep(0,N-num.corr[j]))
    true.signif[j]<-1-pbinom(num.corr[j]-1,N,1/3)
    true.result[j]<-ifelse(true.signif[j]>0.05,"cannot reject the null
hypothesis","reject the null hypothesis")
    for (i in 1:length(num.panl))
    {
      poss.comb<-combn(N,num.panl[i])
      result.vec<-apply(poss.comb,2,function(x) 1-pbinom(sum(vec[x])-
1,num.panl[i],1/3))
      n<-length(result.vec)
      error.mat[j,i]<-
      ifelse(true.signif[j]>signif,(sum(result.vec<=signif))/n,(sum(result.v
ec>signif))/n)
    }
  }
  win.graph()
  plot(num.panl,error.mat[j,],xlab="number of panellists", ylab="error
rate",main=paste(num.corr[j],"successes",sep=" "),type="b")
}
```